EXAMINING THE GENERALIZED WARING MODEL FOR THE ANALYSIS OF TRAFFIC CRASHES

A Dissertation

by

YICHUAN PENG

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	Quadrifoglio Luca
	Yunlong Zhang
	Jeffrey D. Hart
Head of Department	John Niedzwecki

May 2013

Major Subject: Civil Engineering

ABSTRACT

As one of the major data analysis methods, statistical models play an important role in traffic safety analysis. A common situation associated with crash data is the phenomenon known as overdispersion which has been discussed and investigated frequently in recent years. As such, researchers have proposed several models, such as the Poisson Gamma (PG) or Negative Binomial (NB), the Poisson-lognormal, or the Poisson-Weibull, to handle the overdispersion. Unfortunately, very few models have been proposed for specifically analyzing the sources of dispersions in the data. Better understanding of sources of variation and overdispersion could help in managing safety, such as establishing relationships and applying appropriate treatments or countermeasures, more efficiently.

Given the limitations of existing models for exploring the source of overdispersion of crash data, this research examined a new model function that could be applied to explore sources of extra variability through the use of the Generalized Waring (GW) models. This model, which was recently introduced by statisticians, divides the observed variability into three components: randomness, internal differences between road segments or intersections, and the variances caused by other external factors that have not been included as covariates in the model. To evaluate these models, GW models were examined using both simulated and empirical crash datasets, and the results were compared to the most commonly used NB model and the recently developed NB-Lindley models. For model parameter estimation, both the maximum likelihood method and a Bayesian approach were adopted for better comparison.

A simulation study was used to show the better performance of this model compared to NB model for overdispersed data, and then an application in the empirical crash data illustrates its capability of modeling data sets with great accuracy and exploring the source of overdispersion.

The performances of hotspot identification for these two kinds of models (i.e., GW models and NB models) were also examined and compared based on the estimated models from the empirical dataset. Finally, bias properties related to the choice of prior distributions for parameters in GW model were examined by using a simulation study. In addition, the suggestions on the choice of minimum sample size and priors were presented for different kinds of datasets.

ACKNOWLEDGEMENTS

At First, I would like to express my deepest appreciation to my advisor, Dr. Dominique Lord for his constant encouragement and guidance throughout the research. The work should not be completed successfully without his continuous support.

I also would like to express my appreciation to my committee members, Dr. Yunlong Zhang, Dr. Luca Quadrifoglio and Dr. Jeffrey D Hart for their suggestions on this dissertation. Special thanks are given to Dr. Yunlong Zhang for his hearty support during these four years.

I also want to extend my thanks to Texas Transportation Institute (TTI) for providing funding and invaluable practical experience through my PhD study period. I have accumulated much useful and valuable analyzing and modeling skills during the research period. I wish to especially express my gratitude to Dr. Kay Fitzpatrick and Marcus for their advice and help for my research at TTI.

I also wish to give my appreciation to my colleagues and friends, including Srinivas Geedipally, Fan Ye, Hancheng Ge, Chao Huang, Wen Wang, Xiaosi, Lu Wei for their help and healthy discussions during my study.

I would like to express my deep gratitude to my parents for their continuous encouragement and attention. Last but not least, I am grateful to one of my special friends, Zhu Hui, who always encourages me to be optimistic about my future.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Traffic safety is a very important issue all over the world. In 2008, there were more than 37,000 people lost their lives in motor vehicle crashes (NHTSA, 2010). In recent years, this number of traffic fatalities significantly fell to a yearly average of 33,000 (NHTSA, 2010). Unfortunately, this is still a very large number. According to a report published by World Health Organization (WHO), more than a million people lose their lives on the world's roads each year (WHO, 2012). It is also estimated that nearly 50 million people are injured in road traffic crashes around the world each year and traffic crashes are the most important reason of death among teenagers. It has been predicted that road death will probably rise to the fifth leading cause of death by 2030 if without effective countermeasures taken (WHO, 2009). In addition, traffic crashes cause significant economic and social costs to society. The economic cost of road crashes and injuries composes around 1% of gross national product (GNP) in low-income countries, 1.5% in middle-income countries and 2% in high-income countries (WHO, 2009). The global cost related to traffic crashes is estimated to be US$518 billion per year (WHO, 2009). According to the report released from America Automobile Association (AAA, 2011), the costs caused by traffic crashes are more than three times greater than those associated with congestion. People coming from low-income or high-income families are all seriously affected by road traffic crashes. Thus, traffic safety has been a major concern for many government agencies and private organizations.

As one of the major data analysis methods, statistical models play an important role in traffic safety analysis (Geedipally and Lord, 2010). They can be used to explain random variations of accidents across sites based on the available information, such as traffic flow and other road geometric variables. They also can be used to investigate specific or significant effects of the variables on the risk of the collision. In addition, the number of crashes or crash severity can usually be predicted accurately by developing statistical models.

The most commonly used crash count models in traffic crash analysis are the Poisson and Poisson-gamma (PG), also can be called the Negative Binomial (NB) model. Although the performance of Poisson and NB models in modeling crashes is great, there are some limitations for them (Oh et al., 2006). These limitations include the biased goodness-of-fit (GOF) statistics, improper estimation of dispersion parameter, and biased parameter estimates when the crash data are characterized by low sample mean (LSM) and small sample size (SSS) (Lord, 2006). The other important limitation associated with NB model is the assumption of the probability density function (PDF) when the data exhibit underdispersion characteristic, which indicates the mean is greater than the variance. Moreover, the unobserved overdispersion in traffic crash data may be due to internal variances across individuals and external factors that have not been included as covariates in the model. In the NB regression model, both sources of variation in the occurrence rate are jointly considered by means of the gamma distribution. To overcome some of the difficulties described above, researchers have proposed the use of zero-inflated models (Shankar et al., 1997 and 2003) which work as a dual state process, one of which is assumed by having a crash mean equal to zero (Warton, 2005; Lord, et al., 2005 & 2007; Wedagama et al., 2006). The assumption of the dual state process may not be right for analyzing crash data because absolutely safe roadway segments do not exist. Many new methods have been introduced in traffic safety research recently, such as the Beta-binomial model (Tong and Lord, 2007), neural and Bayesian neural network models (Xie et al., 2008), Support Vector Machine (SVM) models (Li et al., 2008), Conway-Maxwell Poisson models (Geddipally, 2008), finite mixture models (Park, 2010), and the NB Lindley generalized linear model (Geedipally and Lord, 2012). Some of them have already represented better performance compared to the traditional NB model in some cases. But the traditional NB model is still the most commonly used model for analyzing traffic crashes. Given the limitations of the NB model mentioned above, it is meaningful to investigate whether better count data models existed for modeling motor vehicle crashes.

This research aims to evaluating the application of the Generalized Waring (GW) model for analyzing crash data. The GW model was recently introduced by Rodriguez-Avi et al. (2009). More specifically, the dissertation expands on the work of Rodriguez-Avi et al. (2009) and evaluates the performance of GW distribution for various datasets with different sample means, sample size and levels of dispersion. Another major objective of this research is to investigate the application of the GW model for analyzing traffic crashes by comparing the performance of the GW model with the standard NB model and the NB Lindley (NB-L) model recently proposed by Lord and Geedipally (2011) and Geedipally et al. (2012). Furthermore, this research will examine the performance of GW distribution including bias for a dataset characterized by small sample size and low sample mean. Finally, the research will develop recommendations for implementing the GW model in traffic safety research and will propose some directions for future research.

1.1 Problem Statement

Traffic crashes are usually considered as random events by traffic analyst by assuming a mean crash rate existed for each roadway segment or intersection. One of the most common characteristics of crash data is overdispersion when the variability accounted for by the Poisson assumption is not sufficient.

Traffic safety analysts accordingly have used NB models by adding gamma error distribution on the base of Poisson assumption of the mean of the number of crashes in order to address the extra Poisson variation. Although there are considerable efforts made to improve the performance of the NB model including using a varying dispersion parameter instead of fixed parameter (Hauer, 2001; Miaou and Lord, 2003), there are still some important limitations associated with these models in a crash analysis.

First, NB model cannot represent the nature of the overdispersion very clearly in the data. It only claims that overdispersion has been observed and addressed in this model (Land et al., 1996). It is difficult to determine the specific sources of variances, especially with a fixed dispersion parameter. It is also difficult to find the sources of variances even if a varying dispersion parameter is used in the model because the overdispersion is partly caused by external covariates that hard to be measured or observed.

Second, it has been reported by some researchers that NB regression models have difficulties fitting heavily overdispersed datasets (Stein et al., 1987; Geddipally et al., 2012). Such datasets are usually characterized by heavy long tail (Guo and Trivedi, 2002).

Third, many empirical crash data exhibit more zero observations than would be allowed by a NB regression model, which causes the low mean issue of traffic crash datasets. It is often difficult to collect large amount of data because of limited time and budget. In these cases that the datasets exhibit small sample sizes and low mean values characteristic, the performance of the parameter estimation using NB model can be significantly affected (Lord, 2006 , Park and Lord, 2008).

Given above-mentioned limitations of the NB regression model for dealing with overdispersion of crash data, this research will focus on examining an alternative model formulation that probably can be used for explaining the nature of overdispersion and fitting highly dispersed crash data better.

1.2 Research Objectives

Though considerable efforts have been taken to developing crash count models to analyze the relationship between crash frequency and various factors, it is still meaningful to examine other potential statistical count models that can be used in traffic safety given the limitation of the most commonly used NB model and other models mentioned above. The GW model was first provided by Irwin to theoretically analyze accidents (Irwin, 1968; Xekalaki, 1983). The main advantage of this model over the NB model is that the former provides more specific information about the variance. It can be used to further distinguish the observed overdispersion that is caused by internal factors inherent to each road segment or intersection from those caused by external factors that are hard to be observed or measured and have not been included in the model.

The following objectives will be addressed in this research:

1. Introduce the GW model and examine the performance of GW distribution using simulated datasets with different sample means and sample sizes by investigating the parameter estimation accuracy of the model in each condition.

2. Apply the GW model to analyzing motor vehicle crashes. Compare the performance of GW model with the standard NB model by using simulated data and empirical data by expanding on the work of Rodriguez-Avi et al. (2009). Also, compare the performance of GW with the recently introduced NB-L model.

3. Evaluate the performance of GW distribution including stability and presence of bias for different kinds of datasets.

4. Examine the performance of the proposed GW model in the ability of identification of hotspots and compare the results between the proposed model and the NB model.

5. Develop recommendations for implementing the GW model in traffic safety research and propose directions for future research.

1.3 Dissertation Outline

The outline of this dissertation is as follows:

Chapter II gives an overview about various crash count data models that have been proposed for modeling highway safety. These models include crash count models for both overdispersion and underdispersion. The major benefits and shortcomings of all these models are discussed in the chapter. In addition, two commonly used parameter estimating methods, the maximum likelihood method and the Bayesian estimating method are provided and discussed.

Chapter III introduces the univariate GW distribution and presents the methodology for analyzing overdispersed crash data by using the univariate GW model. Then, the methodology of the two parameter estimating methods applied into the univariate GW model is provided. Finally, the relationship between the univariate GW model and the NB model is discussed.

Chapter IV mainly examines the performance of the GW models using several simulated datasets. Four hypothetical examples are presented with different purposes. The main objective of this chapter is to show the better prediction capabilities of the GW model compared to the NB model for the overdispersed data.

Chapter V applies the GW model to actual vehicle crash data and the results are compared with those from the NB model and the NB-L in various aspects such as goodness-of-fit, source of variance, and parameter interpretation. Analyses are carried out with four empirical crash datasets: one for an intersection crash dataset and the others for segment crash datasets. The segment crash datasets are empirical highly dispersed vehicle crash data. The main objective of this chapter is to show the major benefit of this model in explaining the nature of overdispersion in different segments or intersections.

Chapter VI deals with the application side of the developed model in terms of the identification of hotspots. The comparison of the results between the proposed model and the NB model is also summarized.

Chapter VII will evaluate the performance of GW distribution in terms of bias in Bayesian statistics for different kinds of simulated datasets.

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

CHAPTER II

BACKGROUND

There have been considerable efforts to develop statistical models for analyzing crash data over the last decades. These models are developed for the purpose of addressing three common properties that can be observed in crash data: overdispersion, non-equal variance, and excess zeros (Park, 2010). Among them, overdispersion is the most common problem and has been paid much attention by researchers. Although there have been a large number of models developed by researchers, the essence in sources of variation is still unclear and need to be investigated further.

This chapter generally presents different kinds of crash count models commonly applied in traffic safety analysis. It is divided into eight sections. Section 2.1 describes the most commonly used Poisson and NB crash count models. The dual state hurdle model, finite mixture model and NB-L model used in traffic safety literature are presented in Sections 2.2, 2.3 and 2.4, respectively. Section 2.5 describes the Conway-Maxwell-Poisson model which can address underdispersion. Some other innovative models are also mentioned in Section 2.6. The two most commonly used parameter estimating methods are discussed in Section 2.7. Finally, Section 2.8 provides a summary for the chapter.

2.1 Poisson and NB Crash Count Models

This section is divided into two parts. The first part introduces the Poisson model and the second part discusses the most commonly used NB model.

2.1.1 Poisson Model

The Poisson distribution is considered to be the basic distribution for analyzing traffic crash data (Lord and Mannering, 2010). In the basic form of a Poisson model, the number of crashes per year, y_i for a particular site i is assumed to follow a Poisson distribution with the mean crashes per year λ_i (Lord and Geedipally, 2008):

$$y_i|\lambda_i \sim \text{Poisson}\ (\lambda_i) \tag{2.1}$$

The mean number of the crashes λ_i is commonly specified as the exponential function of the covariates as:

$$\lambda_i = f(X;\beta) \tag{2.2}$$

where:

$f(.)$ is a function of the covariates (X);

β is a vector of regression coefficients;

X is a vector of traffic flow and site specific covariates.

The probability density function (PDF) of the Poisson distribution is given by the following equation:

$$p(y_i \mid \lambda_{\mathrm{i}}) = \frac{\exp(-\lambda i)\lambda_i^{y_i}}{y_i!} \tag{2.3}$$

The mean and variance of the Poisson distribution is given by

$$E(y_i \mid \lambda_{\mathrm{i}}) = Var(y_i \mid \lambda_{\mathrm{i}}) = \lambda_{\mathrm{i}} \tag{2.4}$$

Poisson model is appropriate when the dependent variable is a count number and was widely used in traffic safety area until the NB model was fully developed (as is described below). The advantages of using the Poisson model over those normal linear regression models including the better appropriateness of model assumptions and improvement of goodness of fit (Joshua and Garber, 1990). In spite of these advantages, a disadvantage of this model lies in the limitation that the variance has to be equal to the mean. Overdispersion is a very common phenomenon in traffic crash datasets. It is not suitable in those overdispersed situations because its limitation to capture the overdispersion.

2.1.2 Negative Binomial Model

Negative Binomial (NB) distribution is the most common distribution used in the traffic safety analysis. This distribution is preferred over other mixed-Poisson distributions by including the gamma error distribution into the Poisson distribution. This model accordingly provides an easy way to deal with overdispersion in this way. The NB model has the following model specification (Lord, 2006): the mean of the number of crashes 'Y_i' for a particular i^{th} site is Poisson distributed and independent over all sites and time periods

$$Y_i \mid \lambda_i \sim Po(\lambda_i) \text{ i = 1, 2,…, I} \tag{2.5}$$

The mean of the Poisson is structured as:

$$\lambda_i = f(X;\beta)\exp(e_i) = \mu_i \exp(e_i) \tag{2.6}$$

where:

$f(.)$ is a function of the covariates (X);

β is a vector of unknown coefficients; and,

e_i is the model error independent of all covariates.

It is assumed that $\exp(e_i)$ follows gamma distribution with a mean equal to 1 and a variance δ for all sites.

$$\exp(e_i) = gamma(1/\delta, 1/\delta) \tag{2.7}$$

where:

δ is the dispersion parameter (note: variance function is $Var(Y_i) = \lambda_i + \delta\lambda_i^2$).

Therefore (Lord and Park, 2010),

$$\begin{aligned} p(y_i) &= \int_0^\infty Pois(y_i \mid \mu_1 ex^{e_i}) g(e^{e_i}) de_i \\ &= \int_0^\infty \frac{(\mu_i e^{e_i})^{y_i} e^{-\mu_i e^{e_i}}}{y_i!} \cdot \frac{(\phi)^t}{\tau(\phi)} (e^{e_i})^{\phi-1} e^{-\phi e^{e_i}} de_i \\ &= \frac{(\phi)^\phi (\mu_i)^{y_i}}{y_i! \tau(\phi)} \int_0^\infty e^{e_i(y_i+\phi-1)} e^{-(\mu_i+\phi)e^{e_i}} de_i \end{aligned} \tag{2.8}$$

ϕ is the inverse of dispersion parameter ($\delta = 1/\phi$).

In addition,

$$\frac{(\mu_i + \phi)^{y_i+\phi}}{\tau(y_i + \phi)} \int_0^\infty e^{e_i(y_i+\phi-1)} e^{-(\mu_i+\phi)e^{e_i}} de_i = 1 \tag{2.9}$$

Therefore, the PDF of the Negative Binomial distribution is given by the following equation:

$$f(y_i \mid \lambda_1, \delta) = \frac{\tau(y_i + \delta^{-1})(\lambda_i \delta)^{y_t}}{\tau(\delta^{-1}) y_i!} (1 + \delta\lambda_i)^{-(y_i + \delta^{-1})} \tag{2.10}$$

The mean and variance of the Negative Binomial distribution is shown in the following equation:

$$E(y_i \mid \lambda_1, \delta) = Var(y_i \mid \lambda_1) = \lambda_1 \tag{2.11}$$

$$Var(y_i \mid \lambda_1, \delta) = \lambda_1 + \delta\lambda_i{}^2 \tag{2.12}$$

It is noted that the dispersion parameter δ is commonly used in different areas of traffic safety analyses from the computation of the weight factor using empirical Bayesian (EB) method (Hauer, 1997; Lord and Park, 2008) to the parameter estimation of crash count models (Geedipally and Lord, 2008). The dispersion parameter is usually considered by traffic safety researchers (Mitra and Washington, 2007) as fixed when using the traditional Negative Binomial model in most cases. They (Mitra and Washington, 2007) have also suggested that the varying dispersion parameter may not be necessary especially when the functional form contains many covariates. On the other hand, other researchers suggested that the variance function should use a varying dispersion parameter (Miaou and Lord, 2003) that can be used to address the site specific attributes. A fixed dispersion parameter will be more appropriate if there are no significant covariates causing the systematic dispersion (Lord and Park, 2008).

Although the NB model is the most commonly used crash count model, there are some important limitations. The primary issue when dealing with crash data is the problem associated with the SSS and the LSM biases. In addition, some researchers have indicated that sometimes a negative dispersion parameter is a misspecification of the PDF of crash data exhibiting underdispersion(Clark and Perry, 1989; Saha and Paul, 2005). Therefore, this model has difficulties converging with datasets exhibiting this characteristic. Another important limitation of the NB model lies in its limited ability in identifying sources of variances that come from internal variances and external variances. For the NB model, sources of variation after excluding randomness are considered together by means of a gamma distribution. It cannot be used to distinguish the part of the overdispersion further into internal factors inherent to each road segment or intersection and external factors that have not been included in the model because of difficulty of observation or measurement.

2.2 Dual State Hurdle Model

This section presents a brief introduction of hurdle models and other dual state models, along with their advantages and limitations. As it was mentioned above, many crash data exhibit extraordinary zero observations than would be allowed by the Poisson or NB model. In highway safety literature, some researchers have applied the hurdle models (Son et al., 2009) and zero-inflated regression models for the purpose of accommodating the excess zeros (Shankar et al., 1997; Shankar et al., 2003).

The basic concept of the hurdle model is that it partitions the data generating process into two parts. The first part models the probability that the value below a threshold is observed, and the second part models the probability that values above the threshold are observed. In principle, the threshold could be any value (Cameron and Trivedi, 1998). The general form of a hurdle model is shown as follows:

$$p(y_i) = p_1(0) \qquad \text{if } y_i = 0 \qquad (2.13)$$

$$p(y_i) = \frac{1 - p_1(0)}{1 - p_2(0)} p_2(y_i) \qquad \text{if } y_i \geq 1 \qquad (2.14)$$

The zero-inflated model is a specific type of the hurdle model in which the zero outcomes can arise from one of two processes. The underlying assumption is that zero crash counts are generated by a dual state process: a perfect safe state which is absolutely safe road or a relative dangerous state with a certain mean of number of crashes. Therefore, zeros may come from both states. The binary process is assumed when modeling the unobserved state. The probability density function is shown as follows:

$$p(y_i) = w_i + (1 - w_i) p_2(0) \qquad \text{if } y_i = 0 \qquad (2.15)$$

$$p(yi) = (1 - w_i) p_2(y_i) \qquad \text{if} \quad y_i \geq 1 \tag{2.16}$$

Although the above-described dual state models have provided a better goodness of fit to data as compared to other count models, they have been criticized in modeling vehicle crash data by some researchers (Lord et al. 2005, 2007) because of the unrealistic underlying assumption that there is a group of sites that never experience crashes. This is unrealistic since a roadway segment or an intersection always has the possibility of generating crashes unless there is no traffic on it.

2.3 Finite Mixture Models

The major advantage of the finite mixture model lies in its more flexible functional form to fit over dispersed data using a combination of several discrete or continuous distributions. This kind of model has been extensively used in many areas (e.g. Ramaswamy et al., 1994; Wang et al., 1998; Guo & Trivedi, 2002)and has been proposed and applied in the traffic safety context recently (Park & Lord, 2010). The general model structure of a finite mixture model can be formulated as follows (Park & Lord, 2010):

$$p(y \mid \theta) = w_1 f_1(y \mid \theta_1) + w_2 f_2(y \mid \theta_2) + \cdots + w_k f_k(y \mid \theta_k) \tag{2.17}$$

where the random vector $y = (y_{1,} y_2 ... y_N)'$ is considered to be composed of several discrete or continuous distributions, $\theta = (\theta_1, \theta_2, \cdots \theta_\text{k})'$, w) represents all parameters for each distribution and $w = (w_{1,} w_2 ... w_k)'$ means a weight vector. The sum of all elements of w is equal to 1. A single density $f_k(. \mid \theta_k)$ represents each distribution and k is the number of distributions. The component distribution is assumed to be a Poisson or NB distribution when the finite mixture model is applied in traffic safety area (Park & Lord, 2010).

The mean and the variance of a finite mixture model are shown in the following equations

$$\mu = E(y \mid \theta) = \sum_{k=1}^{k} \mu_k w_k \tag{2.18}$$

$$\sigma^2 = Var(y \mid \theta) = \sum_{k=1}^{k} (\mu_k^2 + \sigma_k^2) w_k - \mu^2 \tag{2.19}$$

It is assumed that the component moments μ_k and σ_k exist.

The overdispersion of data can be explained in two aspects using this modeling specification. First, it means the overdispersion caused by latent components. The model assumes that there is more than one component in the data set. Second, the overdispersion within each component also can be accounted for by choosing different kinds of distributions for each component. For example, for finite mixture Poisson models and finite mixture NB regression models, the overdispersion in each component is addressed by including Poisson distribution for the number of traffic crashes. The NB distribution is used to explain additional overdispersion non-related to the variables included in the model. In a word, the formulation of the finite mixture model is flexible enough to address both between-component and within-component variations (Park & Lord, 2010).

Although the finite mixture model provides an obvious advantage in fitting overdispersion of population, there are also some limitations to the model. For example, there is still no consensus method for some issues, especially for the label switching problem and how to identify the optimal number of components.

As can be seen from the above mentioned model structure, the traditional Poisson model and NB model are special cases of finite mixture models by setting K =1 in the model, and the hurdle and zero inflated models mentioned in Section 2.2 are also the special cases of the finite mixture model where the optimal number of components are equal to 2.

2.4 Negative Binomial Lindley Model

This Negative Binomial Lindley (NB-L) model was recently proposed by some researchers for analyzing traffic crash data (Lord and Geedipally, 2011). As discussed above, two common characteristics associated with traffic crash datasets are small sample size and low sample mean because they are composed of a large amount of zeros. The performance of traditional Poisson or NB models applied in these highly dispersed datasets are not so good as in other over dispersed datasets while The NB-L model is a good option in these cases.

As it is shown in the following equation, the NB-L distribution is composed of Negative Binomial and Lindley distributions. The PDF of NB-L distribution is calculated by using the following equation.

$$P(Y = y, \mu, \phi, \theta) = \int NB(y; \phi, \varepsilon\mu) Lindley(\varepsilon; \theta) d\varepsilon \qquad (2.20)$$

Here, the parameter μ means the mean of independent variable and ε follows the Lindley distribution (Lord and Geedipally, 2011). The PDF of the Lindley distribution is shown in the following equation:

$$f(X = x; \theta) = \frac{\theta^2}{\theta + 1}(1 + x)e^{-\theta x}; \theta > 0, x > 0 \qquad (2.21)$$

The mean of the Lindley distribution is (Ghitany et al., 2008):

$$E(\varepsilon)=\frac{\theta+2}{\theta(\theta+1)} \tag{2.22}$$

The variance of the Lindley distribution is (Ghitany et al, 2008):

$$E(Y)=\mu\times E(\varepsilon)=e^{\beta_0+\sum_{i=1}^{p}\beta_i X}\ \frac{\theta+2}{\theta(\theta+1)} \tag{2.23}$$

The crash variance is given by the Equation 2.22:

$$Var(Y)=\mu\times\frac{\theta+2}{\theta(\theta+1)}+\mu^2\times\frac{2(\theta+3)}{\theta^2(\theta+1)}\times\frac{1+\phi}{\phi}-(\mu\times\frac{\theta+2}{\theta(\theta+1)})^2 \tag{2.24}$$

According to the results from other researchers, the NB-L fits better than the traditional NB model when datasets are highly dispersed, especially when datasets are characterized by extraordinary zeros or with a heavy tail. When the dispersion becomes smaller, the NB-L model is still able to provide the same performance as the NB model (Lord and Geedipally, 2011). Although the NB-L model provided a better fit, further work including the basic data generating process and whether the model is logically sound still needed to be done on this topic.

2.5 Conway-Maxwell-Poisson Regression Model

The Conway-Maxwell-Poisson (COM-Poisson) distribution, originally proposed by Conway and Maxwell (1962), is a kind of extension of the Poisson distributions. The major advantage of this kind of model lies in its ability to account for both over and underdispersion data. The PDF of the COM-Poisson distribution is shown in the following equation (Geedipally, 2008):

$$\begin{aligned} p(y_i) &= \frac{1}{Z(\lambda,v)}\frac{\lambda^{y_i}}{(y_{i!})^{v}} \\ Z(\lambda,v) &= \sum_{j=0}^{\infty}\frac{\lambda^{j}}{(j!)^{v}} \end{aligned} \tag{2.25}$$

where λ is a centering parameter that is related to the mean of the observations, and v is a shape parameter. This model address underdispersed data overdispersed data and equidispersed data respectively by choosing different values of v.

The mean can be approximately calculated through some different approaches. For example, Shmueli et al. (2005) derived the mean and variance as follows:

$$\begin{aligned} E(y_i) &\approx \lambda^{1/v} + \frac{1}{2v} - \frac{1}{2} \\ Var(y_i) &\approx \frac{1}{v}\lambda^{1/v} \end{aligned} \tag{2.26}$$

This model was applied by Lord et al. (2008) for evaluating vehicle crash data. In this new form, λ is replaced with $\mu = \lambda^{1/v}$ representing a clear centering parameter. Although it is an effective method for analyzing an underdispersed dataset, there are still some limitations in the model, especially when it is used to analyze overdispersed data. The other detailed model characteristics and the performance with application to vehicle crash data can be investigated in Geedipally (2008).

2.6 Other Models

There are several other statistical models used for analyzing vehicle crashes in recent years. These include the multivariate Poisson-lognormal model and the application of Beta-binomial model (Lord et al., 2005; Tong and Lord, 2007). Bayesian network models and support vector machine models have also been used for crash predictions

(Xie et al., 2007; Li et al, 2008). However, the parameter estimation methods for these models are complex and hard to be generalized to all kinds of crash datasets. A two-state Markov switching model which is a kind of finite mixture model has been applied by assuming that there are two states of roadway safety changing over time (Malyshkina et al., 2009). Random parameter models have been initially proposed by Anastasopoulos and Mannering (2009). The performance of this kind of model on goodness of fit is better than models with fixed parameters in most cases. However, the parameter estimation methods are too complex to be applied in every crash dataset.

2.7 Parameter Estimating Methods

In this section, two most commonly used parameter estimating methods are presented and discussed. The first section introduces the maximum likelihood method, and the second discusses the Bayesian method.

2.7.1 Maximum Likelihood Method

This section gives a brief description of the maximum likelihood estimation (MLE). It is the most popular technique and has traditionally been used by many researchers for estimating the model coefficients. The joint probability density function has to be calculated in order to be able to apply MLE parameter estimation method. The PDF that specifies the probability of observing vector(X_1, X_2, X_3… Xn) given the parameter w is shown in the following equation (Myung, 2002):

$$f(\mathrm{x}_{1,}\mathrm{x}_2...\mathrm{x}_\mathrm{n} \mid w) = f(x_1 \mid w).f(x_2 \mid w)...f(x_n \mid w) \tag{2.27}$$

The likelihood function is accordingly defined as follows:

$$L(w \mid \mathrm{x}_{1,}\mathrm{x}_2...\mathrm{x}_\mathrm{n}) = f(\mathrm{x}_{1,}\mathrm{x}_2...\mathrm{x}_\mathrm{n} \mid w) \tag{2.28}$$

Then the maximum value is calculated based on the above equation. When the log-likelihood function is differentiable for being calculated, the maximum likelihood estimation (MLE) of these parameters is obtained by getting the solution of n equations as follows:

$$\frac{\partial}{\partial \theta_i} L(\theta \mid y) = 0, i = 1,...,n) \tag{2.29}$$

The benefits of maximum likelihood estimators are (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook):

1. The parameter estimators are of unbiased minimum variance as the sample size becomes very large. This means that the mean value of each estimated parameter will be theoretically equal to the population value when a large number of samples were taken from the population. Minimum variance leads to the result that the estimator has the smallest variance and the accordingly the confidence interval becomes the narrowest.

2. They have approximately normal distributions and sample variances that can be used to estimate confidence intervals and generate hypothesis tests for the parameters.

3. It is not difficult to be applied even for the large data sets because several different types of statistical software provide existed algorithms for maximum likelihood parameter estimation. According to Fruhwirth-Schnatter (2006), there are a few practical difficulties with the maximum likelihood estimation of regression models. For example, it is difficult to find a global maximum of the likelihood numerically. It is often the case that a local maximum has been found. Therefore, the best way to find the global maximum value is to use many different starting points to compare results. Furthermore, the algorithm is hard to converge when the sample size is too small.

2.7.2 Bayesian Method

The most important difference between the maximum likelihood and Bayesian methods lies in the way they consider the estimated parameter. The maximum likelihood method regards unknown parameter θ as a fixed value, and then estimates the probability distribution of the data being the result of a random event from a fixed parameter space of θ (Gelman et al., 2004). The value of θ is estimated through getting the maximum value of a likelihood function. The uncertainty about parameter estimates is quantified by investigating the difference between different samples. On the other hand, Bayesians regard the unknown parameter θ as a random variable and are interested in the probability distribution of a model parameter (Gelman et al., 2004). This probability distribution is called a posterior distribution, which is a product of a likelihood function and a prior distribution. The hyper-parameters can be assumed either to be known or to be drawn from s second-stage prior distribution. The prior distribution on θ is very important for Bayesians. The uncertainty about parameter estimates is quantified by determining the difference between different prior distributions given the observed data.

The algorithms employed in the Bayesian estimating method are based on the Markov Chain Monte Carlo (MCMC) sampling techniques introduced by Gelman (1984), Tanner and Wong (1987), and Gelfand and Smith (1990). The MCMC sampling methods have their roots in the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953) and they are now become one of the main powerful computational tools in parameter estimation methods.

MCMC sampling methods greatly simplify the complexity of some highly complicated models and make it possible to estimate the corresponding posterior distributions of these parameters with accuracy. Therefore, MCMC methods have made great contribution to the development and propagation of Bayesian theory. Some researchers (Gilks, 1996; Spiegelhalter, 2003) have explained the details of MCMC methods. The more detailed Bayesian method used in the GW model will be discussed in the following chapter.

2.8 Chapter Summary

In this chapter, various crash count models have been described for addressing both overdispersion and underdispersion crash data. The chapter also summarized the two estimating methodologies that can be used for estimating the coefficients of regression models. The different models that have been proposed for addressing overdispersion were the main focus of this chapter since this kind of dataset is the most commonly observed in practice. A brief discussion about benefits and limitations of existing models was presented.

One of the most important shortcomings of the models described in this chapter is related to the fact that the sources of the overdispersed variance in traffic crash data cannot be investigated efficiently. The variance observed in the data can be caused by internal differences across road segments or intersections, and by external unobserved factors that cannot be included in the model. For most models, both sources of variation in the occurrence rate are jointly considered instead of considering them separately. The next chapter describes the characteristics of GW distribution and GW model developed based on the distribution.

CHAPTER III

METHODOLOGY

This chapter explains the methodology of the GW distribution and describes the GW models. It is divided in six sections. Section 3.1 provides a description of the GW distribution. Section 3.2 illustrates the theoretical basis for using the GW model we developed that was based on the distribution. Section 3.3 describes the methodology for analyzing crash data using this model, based on the maximum likelihood framework; Section 3.4 covers the Bayesian method used for this model. Section 3.5 explains the relationship between the GW model and the NB model. Finally, the last section, Section 3.6, provides a summary of the chapter.

3.1 Generalized Waring Distribution Introduction

Classic accident theory (Greenwood and Yule, 1920; Newbold, 1925, 1927) hypothesized that the overall population of road segments are all subjected to the same external factors, but have unequal levels of proneness to crashes. The proneness of a road segment is represented by λ , which is the mean number of crashes incurred under the given exposure conditions. The variable λ is assumed to have a continuous distribution, and the distribution of accidents among individuals with the same levels of proneness is assumed to be Poisson-distributed with exp$\left\{\lambda(\theta-1)\right\}$. We write this as θ =1+α leads to the basis of the traditional NB model (Irwin 1968).

However, we cannot know that all entities (i.e., road segments, intersections, etc.) will be exposed to exactly the same external risk of accident. Differences in exposure to external factors from one section to the next are known as differences in accident liability, as distinguished from constitutional or internal differences which are known as differences in proneness. In practice, the effects of proneness and liability are confounded (that is, they are inseparable) when the Negative Binomial is used. This combination was originally called "susceptibility" by Newbold (1925, 1927). Therefore, Irwin proposed the univariate Generalized Waring distribution (UGWD) in a way that the gamma distribution accounts for liability and assumes a beta distribution to accommodate proneness. Its generating function is shown in the following form (Irwin 1968):

$$\frac{(x-a)_{[k]}}{x_{[k]}} F(a,k,x+k,\theta) \tag{3.1}$$

where F is the hypergeometric series, θ is the generating symbol, and $x_{[k]}$ denotes x(x+ 1) ... (x+k-1). However, it is generally more convenient to put x - a = p and write the equation in the following form (Irwin 1968):

$$\frac{\Gamma(\rho+a)\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(\rho+a+k)}(1+\frac{ak\theta}{\rho+a+k}+\frac{a(a+1)k(k+1)\theta^2}{(\rho+a+k)(\rho+a+k+1)2!}+\frac{a_{[r]}k_{[r]}\theta^r}{(\rho+a+k)_{[r]}r!} \tag{3.2}$$

Here, ρ >0, a >0, k >0 and ρ need not be an integer; in fact, the distribution is symmetrical in both *a* and *k*.

Originally, Irwin was interested in the distribution because the distribution can account for and the dataset with a very heavy tail by choosing certain parameter. For example, some actual biological distributions usually had exceptionally heavy tails. Most theoretical distributions discussed in the literature at that time were totally inadequate to deal with this situation. The exception was the Yule distribution (Yule, 1924); this distribution is actually the particular case of a GW distribution when *a* = 1 and *k* = 1. Moreover, the GW distribution is so flexible that the tail does not have to be very heavy for all values of all the parameters. The equation can be written as (Irwin 1968):

$$\frac{\Gamma(\rho+a)\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(\rho+a+k)}F\{a,k,\rho+a+k,\theta\}= \\ \frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\int_0^1 \mu^{\rho-1}(1-\mu)^{k-1}\left\{\frac{1}{\mu}-\theta(\frac{1}{\mu}-1)\right\}^{-a} d\mu \tag{3.3}$$

The integral in equation (3.3) is a well-known representation of the hyper-geometric function. The GW distribution can be used separately to allow for proneness and for liability. In the Negative Binomial, which is regarded as the distribution for individuals with a proneness v arising from a Poisson distribution for individuals with a liability ($\lambda|v$) and a fixed proneness v, where ($\lambda|v$); that is, λ for a given v has the usual Pearson Type III (gamma) distribution (Irwin 1968):

$$\frac{1}{\Gamma(a)}e^{-a\lambda/\upsilon}(\frac{a\lambda}{\upsilon})^{a-1}d(\frac{a\lambda}{v}) \tag{3.4}$$

If μ now has the beta distribution (Irwin 1968):

$$\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\mu^{\rho-1}(1-\mu)^{k-1}d\mu \tag{3.5}$$

It follows from Equation 3.3 that the GW distribution will give the final distribution of the accidents.

The beta distribution for µ is a general but plausible assumption based on the literature review (Irwin, 1968).

The total variance of the GW distribution is given by (Irwin 1968):

$$\sigma_t^2 = \frac{ak(\rho+k+1)(\rho+a+1)}{(\rho-1)^2(\rho-2)} \tag{3.6}$$

The variance of the random components is:

$$\sigma_a^2 = \frac{ak}{(\rho-1)} \tag{3.7}$$

The variance of the liability component is:

$$\sigma_b^2 = \frac{ak(k+1)}{(\rho-1)(\rho-2)} \tag{3.8}$$

The variance of the proneness component is:

$$\sigma_c^2 = \frac{a^2k(\rho+k+1)}{(\rho-1)^2(\rho-2)} \tag{3.9}$$

3.2 Theoretical Basis of the GW GLM Model

In this section, this researcher will continue to introduce the theoretical methodology of the GW GLM model which was developed based on the GW distribution. The GLM model was developed by Rodriguez-Avi et al. (2009) by considering covariates as explanatory variables determined by some independent variables in the model. In this research, it will be applied to traffic crash analysis. The dependent variable for the GW model applied here is the mean of the number of crashes in each segment or intersection. Moreover, this model specification assumes that proneness is independent and its distribution is the same for all levels of covariates because it contains all of the conditions inherent to each site (Rodriguez-Avi et al., 2009). Based on this assumption, if X ’= ($X_1,\ldots$ X_p) is the vector of the covariates, let v be the proneness and $\lambda_x \mid v$

the liability for a given proneness; the GW GLM model is generated according to the following steps (Rodriguez-Avi et al., 2009):

1. $(Y \mid X)$, λ_x, v~Poisson(λ_x) (3.10)

2. $\lambda_x \mid v$~ Gamma (a_x, v)$\mapsto$ $Y \mid X$, v~NB(a_x,p) with $p = \frac{1}{1+v}$ (3.11)

3. v~beta (ρ, k), that is

$$f(v) = \frac{\tau(k+\rho)}{\tau(k)\tau(\rho)} v^{k-1}(1+v)^{-(k+\rho)}, v > 0, k, \rho > 0 \tag{3.12}$$

Therefore, the PDF of $Y \mid X$ is:

$$f(\mathrm{Y} \mid \mathrm{x}) = \frac{\tau(a_x+\rho)\tau(k+\rho)}{\tau(a_x+k+\rho)\tau(\rho)} \frac{(a_x)_y (k)_y}{(a_x+k+\rho)_y} \frac{1}{y!}, y = 0,1...,\tag{3.13}$$

leading to a UGWD (a_x, k, ρ), where a_x, k, ρ >0 and:

$$(\alpha)_r = \frac{\tau(\alpha+r)}{\tau(\alpha)} if \alpha > 0 \tag{3.14}$$

4. The equation of log-linearity for the mean according to classical regression models is shown in the following equation:

$$E(Y \mid x) = \mu_x = f(X;\beta) \text{ with } \beta^{'} = (\beta_1,...\beta_p). \tag{3.15}$$

$f(.)$ is a function of the covariates X;

Furthermore, the mean of the UGWD is given by:

$$E(Y \mid x) = \frac{a_x k}{\rho - 1} \tag{3.16}$$

Therefore, ρ must be greater than 1 for the positive $E(Y \mid x)$, and then:

$$a_x = \frac{\mu_x(\rho - 1)}{k} \tag{3.17}$$

At last, for the purpose of guaranteeing the conditions k > 0 and ρ > 1, these parameters are written as the following formulation:

$$k = e^{k_0}, \rho - 1 = e^{\rho_0} \tag{3.18}$$

with k_0, $\rho_0 \in$ R.

The above model is called the GW model. The different parts of variances in this regression model are calculated as (Rodríguez et al., 2009):

$$\begin{aligned}
Var(Y \mid x) &= E(Var(Y \mid x, v)) + Var(E(Y \mid x, v)) \\
&= E(a_x v + a_x v^2) + Var(a_x v) \\
&= E(a_x v) + E(a_x v^2) + Var(a_x v) \\
&= \frac{a_x k}{\rho - 1} + \frac{a_x k(k+1)}{(\rho-1)(\rho-2)} + \frac{a_x^2 k(k+\rho-1)}{(\rho-1)^2(\rho-2)} \\
&= \mu_x + \frac{k+1}{\rho-2}\mu_x + \frac{k+\rho-1}{\rho-2}\frac{\mu_x^2}{k}
\end{aligned} \tag{3.19}$$

For a_x, k >0 and ρ >1

It can be seen from the above equations that there are three parts of the variance, in total, addressed by the GW model. The first part of the variance represents that the variability comes from the randomness coming from the assumed Poisson model. The other two parts are liability and proneness respectively. The detail information about all these three parts was shown in Table 3.1.

Table3.1 Partition of the variance in the GW Model (Rodriguez-Avi et al., 2009)

Source of variability	**Variance**	**Variance rate**
Randomness	μ_x	$\frac{\rho-2}{k+\rho-1}\frac{k}{k+\mu_x}$
Liability	$\frac{k+1}{\rho-2}\mu_x$	$\frac{k+1}{k+\rho-1}\frac{k}{k+\mu_x}$
Proneness	$\frac{k+\rho-1}{\rho-2}\frac{\mu_x^2}{k}$	$\frac{\mu_x}{k+\mu_x}$
Total	$\frac{k+\rho-1}{\rho-2}(\mu_x+\frac{\mu_x^2}{k})$	1

It is necessary to point out that the GW model is significantly better at distinguishing the sources of variance than is GW distribution. It is difficult to identify proneness and liability using GW distribution when an insufficient amount of information is made available by the covariates (Irwin, 1968). This problem is solved by using the GW model with more than one covariate because it renders the parameters a and k no longer interchangeable.

3.3 Maximum Likelihood Parameter Estimating Method for The GW Model

In order to use the method of maximum likelihood for the GW model, one first needs to specify the joint density function for this kind of model. The probability density function of the GW model was defined above in Equation 3.16. Therefore, for certain samples the joint density function of the GW model (or the likelihood function) will be (Rodríguez et al., 2008):

$$\prod\nolimits_{i=1}^{n} f(\mathrm{X_i}\,|\,\mathrm{k},\rho,\mathrm{a_x}) = \frac{\tau(a_x+\rho)\tau(k+\rho)}{\tau(a_x+k+\rho)\tau(\rho)}\frac{(a_x)_x(k)_x}{(a_x+k+\rho)_x}\frac{1}{x_i!}, x=0,1...n \tag{3.20}$$

where *n* is the sample size.

The logarithm of this equation is shown is the following equation:

$$\mathrm{Ln}\prod\nolimits_{i=1}^{n} f(\mathrm{X_i}\,|\,\mathrm{k},\rho,\mathrm{a_x}) = \sum\nolimits_{i=1}^{n}\begin{bmatrix}\ln\tau(a_x+\rho)+\ln\tau(k+\rho)+\ln(a_x)_x+\ln(k)_x \\ -\ln\tau(a_x+k+\rho)-\ln\tau(\rho)-\ln(a_x+k+\rho)_x-\ln x_i!\end{bmatrix} \tag{3.21}$$

Therefore, the maximum likelihood equation is as follows:

$$\begin{gathered}\frac{1}{n}\sum_{i=1}^{n}\psi(1-a_x-x_i)=\psi(1-a_x)+\psi(k+\rho)-\psi(\rho) \\ \frac{1}{n}\sum_{i=1}^{n}\psi(1-k-x_i)=\psi(1-k)+\psi(a_x+\rho)-\psi(\rho) \\ \frac{1}{n}\sum_{i=1}^{n}\psi(1-a_x-k-\rho-x_i)=\psi(1-a_x-k-\rho)+\psi(k+\rho)-\psi(a_x+\rho)-\psi(a_x+k+\rho)-\psi(\rho)\end{gathered} \tag{3.22}$$

Because the log-likelihood function is difficult to solve, the *nlm* and *optim* functions of R were used for the parameter estimation of the GW model (Rodríguez et al, 2009). These two functions were created according to Nelder-Mead, quasi-Newton, and conjugate-gradient algorithms (Team, 2007).

The Nelder–Mead method is a commonly used nonlinear optimization technique, and the quasi-Newton methods are algorithms based on Newton's method for finding the local maxima and minima of functions (John and Kurtis, 2004).

The quasi-Newton and conjugate gradient method are the other two algorithms for the numerical solution of certain linear equations by applying an iterative method and can be applied to equations too complex to be handled by direct methods.

Several initial values were used for the GW model in order to get a global optimum solution for the parameter estimation. The different estimation results became obvious only when these solutions were in the boundary of the parametric space (Rodríguez et al., 2009).

3.4 Bayesian Estimating Method for GW Models

The treatment of over-dispersion is made more explicit by introducing random effects into the Poisson mean (λ_i) in hierarchical Poisson regression models. Depending on the different distributions imposed on the bases of Poisson mean, various mixed Poisson regression models such as Poisson-gamma, Poisson-gamma-Lindey, and Poisson-Gamma-beta can be derived. In this section, we present the cases for Poisson-Gamma-beta within the Bayesian framework.

Within the Bayesian framework of the GW model, as was mentioned in Section 3.2, the mean response for the number of crashes, *y*, has the following hierarchical formulation (Irwin 1968):

1.$(Y \mid x) \sim Poisson\ (\lambda_x)$ (3.23)

2. $\lambda_x \sim Gamma(a_x,\ v)$ (3.24)

3. $v \sim beta\ (\rho\ , k)$ (3.25)

which leads to the conclusion that the PDF of $Y \mid X$ is:

$$f(\mathrm{y}\,|\,\mathrm{x}) = \frac{\tau(a_x+\rho)\tau(k+\rho)}{\tau(a_x+k+\rho)\tau(\rho)} \frac{(a_x)_y (k)_y}{(a_x+k+\rho)_y} \frac{1}{y!}, y = 0,1..., \tag{3.26}$$

Base on the above mentioned formulation, the Bayesian parameter estimation method can be applied and calculated using WINBUGS. In detail, the number of crashes at each site follows the Poisson distribution, and the site-specific error follows a gamma distribution as a prior distribution. The shape parameter of the gamma distribution follows a beta distribution as a prior distribution. Accordingly, the GW model can be considered a three-hierarchy model involving a standard distribution at all levels. In this study, normal priors (non-informative) for β and gamma priors (non-informative and vague) for ρ and *k* are used.

3.5 The Relationship between GW and NB

The GW converges to the NB in two ways (Rodríguez-Avi et al., 2009):

First, if k, $\rho \to \infty$

$$\begin{aligned} f(y \mid x) &\propto \frac{(a_x)_y}{y!} \frac{(\theta(\rho-1))_y}{(a_x+(1+\theta)\rho-\theta)} \\ &= \frac{(a_x)_y}{y!} \frac{\theta^y \rho^y + o(\rho^{(r-1)})}{(1+\theta)^y \rho^y + o(\rho^{(r-1)})} \to \frac{(a_x)_y}{y!} (\frac{\theta}{1+\theta})^y \end{aligned} \tag{3.27}$$

That is, the kernel of the NB $(a_x, \frac{\theta}{1+\theta})$ density, where $\theta = \frac{k}{\rho-1}$. Let us observe that $\mu_x = a_x\theta$ and $Var(Y \mid x) = \mu_x(1+\theta)$. Therefore, the variance is a linear function of the mean and the Negative Binomial I model is obtained. As discussed in Cameron and Trivedi (1998), the functional form of variance for NB I is slightly different than NB II because the variance is a linear function of the mean. This kind of model usually is less flexible in capturing the variance and is not used very often by traffic safety analysts; the NB II is the parameterization of Negative Binomial model most commonly used in count data analysis.

Similarly, if $\rho \to \infty$ and $\theta_x = \mu_x / k$ is bounded, then $a_x \to \infty$ with the same order of convergence, and:

$$\begin{aligned} f(y \mid x) &\propto \frac{(k)_y}{y!} \frac{(\theta_x(\rho-1))_y}{(k+(1+\theta_x)\rho-\theta_x)} \\ &= \frac{(k)_y}{y!} \frac{\theta_x{}^y \rho^y + o(\rho^{(r-1)})}{(1+\theta_x)^y \rho^y + o(\rho^{(r-1)})} \to \frac{(k)_y}{y!} (\frac{\theta_x}{1+\theta_x})^y \end{aligned} \tag{3.28}$$

That is, the kernel of the NB (k, $\frac{1}{1+\theta_x}$) density. $Var(Y \mid x) = \mu_x (1+\frac{\mu_x}{k})$ is calculated based on the PDF, therefore a Negative Binomial II model is obtained(see Cameron and Trivedi, 1998), which is the most commonly used model in traffic safety. From these results, it can be concluded that two different parameterizations of the NB model are nested in the GW model.

3.6 Chapter Summary

This chapter first described the methodology of the univariate GW distribution and the corresponding GW model based on that distribution. The GW distribution was developed such that the gamma distribution models were one of these sources of variation (liability) and they served to introduce a beta distribution for proneness. Therefore, the GW distribution is more flexible than the NB distribution since it can account for the variance by separating it into two components: liability and proneness. The next chapter presents the results of the simulation analysis.

CHAPTER IV

SIMULATED ANALYSIS

In this chapter, the performance of GW model using several simulated datasets is examined. The main objective of this chapter is to examine the performance of the GW model specification in describing the count data which exhibits over-dispersion compared to other count models recently used and developed by other researchers. The maximum likelihood method is used to analyze all the four scenarios of simulated datasets. It is effective to illustrate some theoretical performances of the GW models by using simulated datasets.

Four scenarios are presented in this chapter. The first scenario, described in Section 4.1, is used to compare the performance of the GW and NB models when the underlying distribution is generated from a GW distribution. The effects of sample mean and sample size on the goodness of fit are also examined. The second scenario shown in Section 4.2 is used to illustrate the performance of the GW model specifications when the data are generated from a NB distribution. The effects of sample size and degree of dispersion on the goodness of fit are also examined. The performance of the GW and NB models is also compared in further when the data are generated from a two-component finite mixture of Poisson distribution and a two-component finite mixture of NB distribution. The results are presented in Section 4.3 and Section 4.4. The last section provides a summary of the analysis described in this chapter.

4.1 Scenario 1

This section describes the simulation results for the first scenario. The objective of this scenario is to demonstrate the poor performance of the traditional NB model to properly account for the data generated by a GW distribution. To expand on the work of

Rodriguez-Avi et al. (2009) and simplify the analysis with previous work, the same simulation protocol used by these authors were utilized in this part of the research.

4.1.1 Data Generation Method

For generating GW random variables, two independent covariates $X_i = (X_{1i}, X_{2i})$ were used in the link function, which were generated from the standard normal distribution. The independent variable was then constructed from the two covariates by assuming a log-linear relationship using assigned regression coefficients $\beta_i = (\beta_0, \beta_1, \beta_2)$

The random variables (y) were simulated in the following steps:

Step 1: Set the sample size, β and the parameters of GW distribution to the required values.

The datasets were generated based on three different mean values by selecting the following values of the parameter:

Low mean value: k =2.5, ρ =3.5, β_0 =0.5, β_1= 0.5, β_2=-0.5.

Moderate mean value: k =2.5, ρ =3.5, β_0 =1, β_1= 0.5, β_2=-0.5.

High mean value: k =2.5, ρ =3.5, β_0 =1.5, β_1= 0.5, β_2=-0.5.

The value of k and ρ were assigned based on the consideration that the values of liability and proneness are easy to be observed. Sample sizes equal to 100, 500 and 1,000 were analyzed separately for this simulation.

Step 2: Generate two covariates (X_{1i} , X_{2i}) from the standard normal distribution.

Step 3: Calculate the parameter μ_i using the log-linear function, which is used in normal generalized linear models (Hilbe, 2011). The functional form is as follows: $\mu_i = e^{\beta_0 + x'\beta} = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}$

Step 4: Generate the count variable y_i from the univariate GW distribution UGWD (a_i , k , ρ) where $a_i = \mu_i (\rho - 1) / k$ by using the *rghyper* function in Suppdist R package (Rodríguez-Avi et al., 2009).

Step 5: The simulated datasets were then fitted by using both GW and NB models.

Step 6: Repeat Steps 1 through 5 50 times and compare the average value of simulated results with the theoretical values and assess the general performance of the models.

4.1.2 Modeling Results

As has been mentioned in the previous chapter, the likelihood function for the GW model is given by Equation 3.16 and can be solved using non-linear optimization techniques. This section shows the modeling results when the GW and NB models were fitted with the simulated data generated from GW distribution. Tables 4.1-4.3 summarize the parameter estimation results together with bias and Root Mean Square Error for the datasets with low sample mean values (Oh et al., 2003).

It can be seen from the following tables that the sample size has a significant effect on the estimation process. The values of estimated parameter are closer to the true value when sample size increases from 100 to 1000. The estimates of k , ρ , β_0 , β_1 and β_2 have low bias and RMSE for GW model when sample size above 500 while the bias and RMSE is much higher when sample size is 100. In other words, the GW model was able to reproduce the "true" parameters for the low sample mean over-dispersed data generated by GW with a relative large sample size.

In addition, the better performance of GW model compared to NB model for low mean over-dispersed datasets generated by GW distribution can be seen from Table 4.1 that the value of AIC and BIC are constantly smaller for GW than NB model. It can also be seen from the Table 4.2 and Table 4.3 that the values of estimated parameter are closer to the true value for GW compared with NB for different sample size because the estimates of β_0, β_1 and β_2 have lower bias and RMSE for GW model.

Table 4.1. Parameter estimation results for both models using data generated by GW with low means

	True values	**N=100**		**N=500**		**N=1000**	
		NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	-0.5	-0.36 (0.32)	-0.39 (0.15)	-0.58 (0.22)	-0.54 (0.15)	-0.58 (0.09)	-0.47 (0.05)
β_1	0.5	0.65 (0.33)	0.62 (0.14)	0.46 (0.21)	0.53 (0.16)	0.47 (0.10)	0.49 (0.06)
β_2	-0.5	-0.34 (0.31)	-0.39 (0.13)	-0.54 (0.21)	-0.52 (0.16)	-0.56 (0.11)	-0.48 (0.05)
ρ	3.5	-	4.31	-	3.94	-	3.82
k	2.5	-	3.51	-	2.68	-	2.59
ϕ	-	0.47 (0.18)	-	0.46 (0.11)	-	0.44 (0.08)	-
-2LL	The smaller the better	245.1 (6.97)	**224.5** (6.02)	1088.5 (26.38)	**1054.3** (25.81)	2160 (45.91)	**2120** (42.51)
AIC	”	253.1 (6.97)	**234.5** (6.02)	1096.5 (26.38)	**1064.3** (25.81)	2168 (45.91)	**2130.0** (42.51)
BIC	”	263.5 (6.97)	**247.5** (6.02)	1113.3 (26.38)	**1085.3** (25.81)	2187.6 (45.91)	**2155.2** (42.51)

NOTE: ( ) indicate the standard error of the estimate
bold characters represent best values and all the following tables are the same
Sample mean=0.76 for sample size N=100
Sample mean=0.71 for sample size N=500
Sample mean=0.65 for sample size N=1000

Table 4.2 Bias summaries using data generated by GW with low means

Estimated parameters	N=100		N=500		N=1000	
	NB	GW	NB	GW	NB	GW
β_0	0.14	**0.11**	-0.08	**-0.07**	-0.08	**0.03**
β_1	0.15	**0.12**	-0.04	**0.03**	-0.03	**-0.01**
β_2	0.16	**0.11**	-0.04	**-0.02**	-0.06	**0.02**

Table 4.3 RMSE summaries using data generated by GW with low means

Estimated parameters	N=100		N=500		N=1000	
	NB	GW	NB	GW	NB	GW
β_0	0.44	**0.33**	0.33	**0.17**	0.12	**0.06**
β_1	0.46	**0.27**	0.31	**0.16**	0.10	**0.06**
β_2	0.44	**0.25**	0.31	**0.16**	0.13	**0.05**

Tables 4.4-4.6 summarize the parameter estimation results together with bias and RMSE for the moderate sample mean values. It can be seen from the following tables that the sample size still has a significant effect on the estimation process, although the effect is not as significant as for dataset with a lower sample mean. The values of estimated parameter are closer to the true value when sample size increases from 100 to 1000. The estimates of k, ρ, β_0, β_1 and β_2 have low bias and RMSE for GW model when sample size above 500 and the estimates of these parameters have obviously lower bias and RMSE than those datasets with low sample mean for the small sample size 100. In other words, the GW model is able to better reproduce the theoretical parameters for the moderate sample mean over-dispersed data generated by GW compared to datasets with low sample mean.

Similar as the modeling results from datasets with low sample mean, the better performance of GW model compared to NB model for moderate mean over-dispersed datasets generated by GW distribution can be seen from Table 4.4 that the value of AIC and BIC are constantly smaller for GW than NB model. It can also be seen from the Table 4.5 and Table 4.6 that the values of estimated parameter are closer to the true value for GW compared with NB for different sample size because the estimates of β_0, β_1 and β_2 have lower bias and RMSE for GW model.

Table 4.4. Parameter estimation results for both models using data generated by GW with moderate means

	True values	N=100		N=500		N=1000	
		NB	GW	NB	GW	NB	GW
β_0	1	1.07 (0.41)	0.95 (0.20)	1.04 (0.32)	1.01 (0.13)	1.03 (0.08)	0.99 (0.03)
β_1	0.5	0.39 (0.41)	0.46 (0.21)	0.55 (0.31)	0.52 (0.14)	0.45 (0.08)	0.48 (0.02)
β_2	-0.5	-0.57 (0.40)	-0.53 (0.21)	-0.47 (0. 31)	-0.52 (0.13)	-0.52 (0.08)	-0.49 (0.02)
ρ	3.5	-	3.82	-	3.68	-	3.47
k	2.5	-	1.63	-	2.71	-	2.68
ϕ	-	0.80 (0.18)	-	0.81 (0.09)	-	0.82 (0.07)	-
-2LL	The smaller the better	456.1 (11.85)	**415.4** (10.53)	2174 (46.32)	**2089.1** (42.11)	4449 (89.21)	**4382.5** (82.35)
AIC	”	464.1 (11.85)	**425.4** (10.53)	2182 (46.32)	**2099.1** (42.11)	4457(89.21)	**4392.5** (82.35)
BIC	”	465.3 (11.85)	**438.4** (10.53)	2186.4 (46.32)	**2120.3** (42.11)	4462.8(89.21)	**4417.0** (82.35)

Table 4.5 Bias summaries using data generated by GW with moderate means

Estimated parameters	N=100		N=500		N=1000	
	NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	0.07	**-0.05**	0.04	**0.01**	0.03	**-0.01**
β_1	-0.11	**-0.04**	0.05	**0.02**	-0.05	**-0.02**
β_2	-0.07	**-0.03**	0.03	**-0.02**	-0.02	**0.01**

Table 4.6 RMSE summaries using data generated by GW with moderate means

Estimated parameters	N=100		N=500		N=1000	
	NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	0.42	**0.21**	0.32	**0.13**	0.09	**0.03**
β_1	0.42	**0.21**	0.31	**0.14**	0.09	**0.03**
β_2	0.41	**0.21**	0.31	**0.13**	0.08	**0.02**

NOTE: ( ) indicate the standard error of the estimate
Sample mean=3.2 for sample size N=100
Sample mean=3.0 for sample size N=500
Sample mean=2.9 for sample size N=1000

Tables 4.7-4.9 summarize the parameter estimation results together with bias and RMSE for the high sample mean values. It can be seen from the following tables that the effect of sample size is not as significant as for datasets with low and moderate sample mean. The values of estimated parameter are always close to the true value when sample size increases from 100 to 1000. The estimates of k, ρ, β_0, β_1 and β_2 always have low bias and RMSE for GW model for datasets with different kinds of sample size and the estimates of these parameters have lower bias and RMSE than those datasets with low and moderate sample mean, especially for the small sample size 100. In other words, the GW model was able to replicate assigned parameters best for the high sample mean over-dispersed data generated by GW compared to datasets with low sample mean and moderate mean.

Similar as the modeling results from datasets with low sample mean and moderate mean, the better performance of GW model compared to NB model for high mean over-dispersed datasets generated by GW distribution can be seen from Table 4.7 to Table 4.9.

Table 4.7 Parameter estimation results for both models using data generated by GW with high means

	True values	**N=100**		**N=500**		**N=1000**	
		NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	1.5	1.52 (0.35)	1.48 (0.18)	1.49 (0.27)	1.5 (0.09)	1.50 (0.05)	1.50 (0.01)
β_1	0.5	0.47 (0.34)	0.52 (0.18)	0.51 (0.26)	0.49 (0.09)	0.47 (0.05)	0.49 (0.01)
β_2	-0.5	-0.54 (0.35)	-0.49 (0.17)	-0.46 (0.26)	-0.46 (0.09)	-0.48 (0.05)	-0.49 (0.01)
ρ	3.5	-	3.42	-	3.58	-	3.55
k	2.5	-	3.4	-	3.1	-	2.8
ϕ	-	0.95 (0.18)	-	1.02 (0.08)	-	0.96 (0.05)	-
-2LL	The smaller the better	531.1 (13.25)	**497.2** (12.15)	2641.6 (49.37)	**2592.3** (47.87)	5271.1 (97.62)	**5160.6** (94.36)
AIC	”	539.1 (13.25)	**507.2** (12.15)	2649.6 (49.37)	**2602.3** (47.87)	5279.1 (97.62)	**5170.6** (94.36)
BIC	”	549.5 (13.25)	**520.2** (12.15)	2666.4 (49.37)	**2623.4** (47.87)	5298.7 (97.62)	**5195.1** (94.36)

Sample mean=5.4 for sample size N=100
Sample mean=5.2 for sample size N=500
Sample mean=5.5 for sample size N=1000

Table 4.8 Bias summaries using data generated by GW with high means

Estimated parameters	**N=100**		**N=500**		**N=1000**	
	NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	0.02	-0.02	-0.01	**0**	0	**0**
β_1	-0.03	**0.02**	0.01	-0.01	-0.03	-0.01
β_2	-0.04	**0.01**	0.04	**0.04**	0.02	**0.01**

Table 4.9 RMSE summaries using data generated by GW with high means

Estimated parameters	**N=100**		**N=500**		**N=1000**	
	NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	0.35	**0.18**	0.27	**0.09**	0.05	**0.01**
β_1	0.34	**0.18**	0.26	**0.09**	0.06	**0.01**
β_2	0.35	**0.17**	0.26	**0.10**	0.05	**0.01**

In conclusion, GW model performs better when the sample mean increases and sample size has a significant effect on the estimation process. The values of estimated parameter are closer to the true value when sample size increases. The estimates of β_0, β_1 and β_2 always have low bias and low standard errors for GW.

It also can be seen from the above tables that GW significantly fits better than NB for the over-dispersed data generated by GW because the value of AIC and BIC is constantly smaller for GW than NB and the bias and RMSE of estimated parameter are lower for GW compared with NB for different sample size and sample mean.

On the other hand, the plots in Figures 4.1-4.3 visualize the goodness-of-fit comparison between the generated (or observed) frequencies and the predicted frequencies from two different kinds of models. It clearly shows that the GW model provides better result than the NB model.

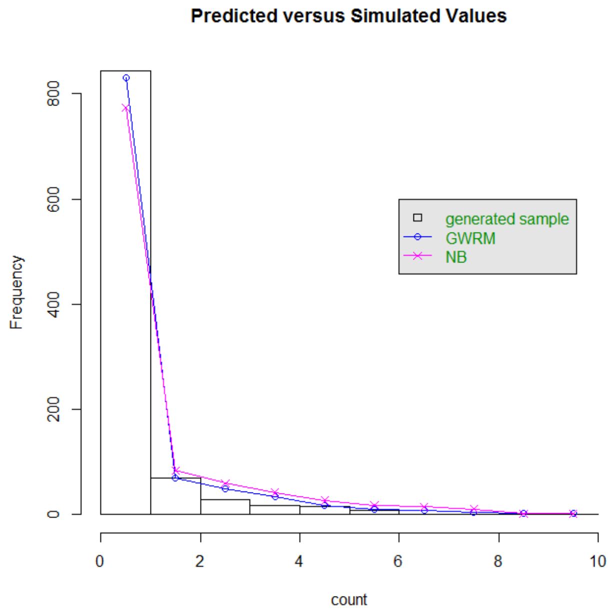

Figure 4.1 Predicted vesus simulated values for GW data of low mean

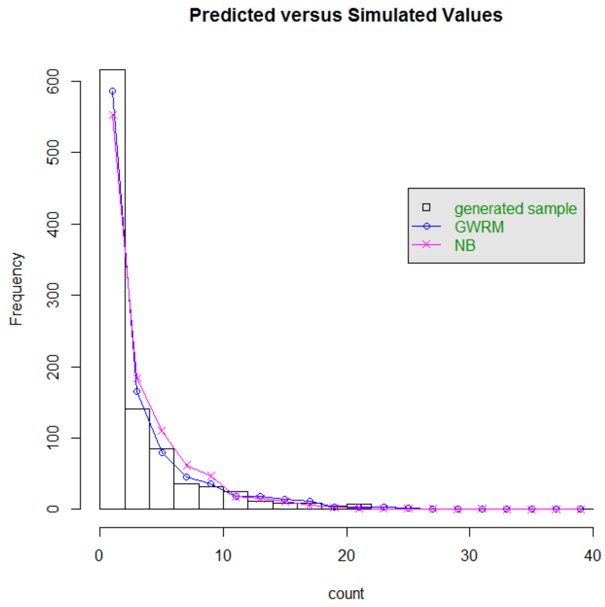

Figure 4. 2 Predicted vesus simulated values for GW data of moderate mean

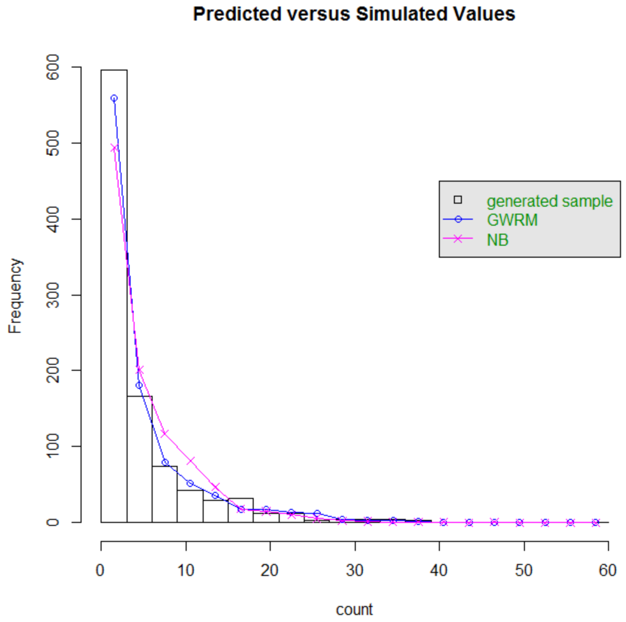

Figure 4. 3 Predicted vesus simulated values for GW data of high mean

4.2 Scenario 2

The objective of this scenario is to examine how well a GW regression model can approximate (or replicate) the data when they are originally generated by a NB distribution. In this scenario, the effect of dispersion parameter of Negative Binomial distribution is also analyzed.

4.2.1 Data Generation Method

For generating NB random variables, a similar simulation protocol as the one described for the first scenario was used. That is, the count was produced using a two-covariate functional form model, where the Poisson mean was assumed to be gamma distributed.

The random variables (y) were simulated in the following steps:

Step 1: Set the sample size, β and the parameters of Negative Binomial distribution to the required values.

In detail, the datasets were generated based on three scenarios of different dispersion values by selecting the following values of the parameter:

Low dispersion value: ϕ=2, β_0 =1, β_1= 0.5, β_2=-0.5.
Moderate dispersion value: ϕ=1, β_0=1, β_1= 0.5, β_2=-0.5.
High dispersion value: ϕ=0.5, β_0=1, β_1= 0.5, β_2=-0.5.

ϕ is the inverse of dispersion parameter. Sample size 100,500 and 1000 were analyzed separately for this simulation.

Step 2: Generate two covariates (X_{1i}, X_{2i}) from the standard normal distribution.

Step 3: Calculate the parameter by using assumed log-linear function, which was commonly encountered in highway crash analysis $\mu_i = e^{\beta_0 + x'\beta} = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}$.

Step 4: Generate the count variable Y_i from the PG distribution (μ_i, ϕ). This simulation protocol will provide similar values as if the counts were produced from the PG protocol.

Step 5: The simulated datasets were then fitted by using both GW and NB models.

Step 6: Repeat Steps 1 through 5 50 times and compare the average value of simulated results with the theoretical values and assess the general performance of the models.

4.2.2 Modeling Results

This section documents the modeling results when the GW and NB models were fitted with the simulated data generated from NB distribution. Tables 4.10-4.12 summarize the parameter estimation results together with bias and RMSE for the highly overdispersed datasets.

It can be seen from the following tables that the sample size still has a significant effect on the estimation process, although the effect is not as significant as in datasets generated from GW distribution. The values of estimated parameter are closer to the true value when sample size increases from 100 to 1000. The estimates of β_0, β_1 and β_2 have low bias and RMSE for both GW and NB model when sample size above 500 and the value of AIC and BIC for both models are almost the same for all the highly dispersed datasets. In other words, the GW model was able to reproduce the theoretical parameters and provide the same performance as NB model for the highly over-dispersed data generated by NB distribution with a relative large sample size.

Table 4.10 Parameter estimation results for both models using data generated by NB with high-dispersed datasets (Φ=0.5)

	True values	N=100		N=500		N=1000	
		NB	GW	NB	GW	NB	GW
β_0	1	0.91 (0.18)	0.91 (0.20)	1.02 (0.07)	1.02 (0.06)	1.01 (0.04)	1.01 (0.04)
β_1	0.5	0.47 (0.16)	0.47 (0.14)	0.52 (0.07)	0.52 (0.06)	0.51 (0.04)	0.50 (0.04)
β_2	-0.5	-0.55 (0.18)	-0.56 (0.15)	-0.51 (0.07)	-0.51 (0.07)	-0.48 (0.04)	-0.48 (0.04)
ρ	-	-	31.5	-	33.4		34.2
k	-	-	0.57	-	0.61		0.64
ϕ	0.5	0.56(0.17)	-	0.52(0.11)	-	0.49(0.06)	
-2LL	The smaller the better	438.1 (11.68)	**434.6** (11.67)	2147.6 (52.14)	**2147.5** (52.14)	4296.1 (98.32)	**4295.4** (98.32)
AIC	”	446.1 (11.68)	**444.6** (11.67)	**2155.6** (52.14)	2157.5 (52.14)	**4304.1** (98.32)	4305.4 (98.32)
BIC	”	**456.5** (11.68)	457.6 (11.67)	**2172.5** (52.14)	2178.5 (52.14)	**4323.7** (98.32)	4329.9 (98.32)

NOTE: ( ) indicate the standard error of the estimate

Table 4. 11 Bias summaries for data generated by NB with high-dispersed datasets

Estimated parameters	N=100		N=500		N=1000	
	NB	GW	NB	GW	NB	GW
β_0	-0.09	-0.09	0.02	0.02	0.01	0.01
β_1	-0.03	-0.03	0.02	0.02	0.01	0
β_2	-0.05	-0.06	-0.01	-0.01	0.02	0.02

Table 4. 12 RMSE summaries using data generated by NB with high-dispersed datasets

Estimated parameters	N=100		N=500		N=1000	
	NB	GW	NB	GW	NB	GW
β_0	0.20	0.22	0.07	0.06	0.04	0.04
β_1	0.16	0.14	0.07	0.06	0.04	0.04
β_2	0.19	0.16	0.07	0.07	0.04	0.04

Tables 4.13-4.15 summarize the parameter estimation results together with bias and RMSE for the moderate-dispersed datasets. It can be seen from the following tables that the sample size has a less effect on the estimation process compared to highly overdispersed datasets generated from NB distribution. The values of estimated parameter are closer to the true value when sample size increases from 100 to 1000 but the difference is not very obvious. The estimates of β_0, β_1 and β_2 have very low bias and RMSE for both GW and NB model when sample size above 500 and the value of AIC and BIC for both models are almost the same for all the moderate dispersed datasets. It can be concluded that the GW model was able to reproduce the "true" parameters for a relative large sample size of moderate dispersed data and provide the same performance as NB model for the moderate dispersed data generated by NB distribution.

Table 4. 13 Parameter estimation results for both models using data generated by NB with moderate-dispersed datasets (Φ=1)

	True value	**N=100**		**N=500**		**N=1000**	
		NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	1	0.94 (0.11)	0.94 (0.10)	0.99 (0.06)	0.99 (0.06)	1.01 (0.04)	1.01 (0.04)
β_1	0.5	0.53 (0.11)	0.54 (0.12)	0.48 (0.06)	0.48 (0.06)	0.51 (0.04)	0.51 (0.04)
β_2	-0.5	-0.48 (0.12)	-0.48 (0.13)	-0.48 (0.06)	-0.48 (0.06)	-0.49 (0.04)	-0.49 (0.04)
ρ	-	-	64.2	-	73.8		72.3
k	-	-	1.1	-	1.6		1.4
ϕ	1	1.1(0.23)	-	0.93(0.09)	-	0.99(0.06)	
-2LL	The smaller the better	421.7 (11.32)	**421.4** (11.32)	2123.5 (51.24)	**2122.6** (51.23)	4324.1 (96.32)	**4324.0** (96.32)
AIC	”	**429.7** (11.32)	431.4 (11.32)	**2131.5** (51.24)	2132.6 (51.23)	**4332.1** (96.32)	4334.0 (96.32)
BIC	”	**440.1** (11.32)	444.4 (11.32)	**2148.3** (51.24)	2153.6 (51.23)	**4351.7** (96.32)	4358.5 (96.32)

NOTE: ( ) indicate the standard error of the estimate

Table 4. 14 Bias summaries using data generated by NB with moderate-dispersed datasets

Estimated parameters	**N=100**		**N=500**		**N=1000**	
	NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	-0.06	-0.06	-0.01	-0.01	0.01	0.01
β_1	0.03	0.04	-0.02	-0.02	0.01	0.01
β_2	0.02	0.02	0.02	0.02	0.01	0.01

Table 4. 15 RMSE summaries using data generated by NB with moderate-dispersed datasets

Estimated parameters	**N=100**		**N=500**		**N=1000**	
	NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	0.13	0.12	0.06	0.06	0.04	0.04
β_1	0.11	0.13	0.06	0.06	0.04	0.04
β_2	0.12	0.13	0.06	0.06	0.04	0.04

Tables 4.16-4.18 summarize the parameter estimation results together with bias and RMSE for the low-dispersed datasets. It can be seen from the following tables that the effect of sample size is not very obvious on the estimation process. The values of estimated parameter are closer to the true value when sample size increases from 100 to 1000. The estimates of β_0, β_1 and β_2 have low bias and RMSE for both GW and NB model especially when sample size above 500 and the value of AIC and BIC for both models are almost the same for all the low dispersed datasets. It can be concluded that the GW model was able to provide the same performance as NB model for the low dispersed data generated by NB distribution and to reproduce the "true" parameters.

Table 4.16 parameter estimation results for both models using data generated by NB with low-dispersed datasets (Φ=2)

	True values	**N=100**		**N=500**		**N=1000**	
		NB	**GW**	**NB**	**GW**	**NB**	**GW**
β_0	1	0.97 (0.08)	0.97 (0.08)	1.01 (0.05)	1.01 (0.05)	1.00 (0.03)	0.99 (0.03)
β_1	0.5	0.49 (0.08)	0.49 (0.08)	0.49 (0.05)	0.49 (0.05)	0.51 (0.03)	0.51 (0.03)
β_2	-0.5	-0.48 (0.10)	-0.48 (0.10)	-0.51 (0.05)	-0.51 (0.05)	-0.49 (0.03)	-0.49 (0.03)
ρ	-	-	22354	-	23024		21265
k	-	-	1.85	-	1.89		2.14
ϕ	2	1.95 (0.42)	-	1.97 (0.21)	-	2.02 (0.14)	
-2LL	The smaller the better	412.3 (11.16)	**412.2** (11.16)	2181.6 (50.62)	**2181.5** (50.62)	4156.9 (94.69)	**4156.7** (94.69)
AIC	"	**420.3** (11.16)	422.2 (11.16)	**2189.6** (50.62)	2191.5 (50.62)	**4164.9** (94.69)	4166.7 (94.69)
BIC	"	**430.7** (11.16)	435.2 (11.16)	**2206.4** (50.62)	2212.5 (50.62)	**4184.5** (94.69)	4191.2 (94.69)

NOTE: ( ) indicate the standard error of the estimate

Table 4.17 Bias summaries using data generated by NB with low-dispersed datasets

Estimated parameters	N=100		N=500		N=1000	
	NB	GW	NB	GW	NB	GW
β_0	-0.03	-0.03	0.01	0.01	0	-0.01
β_1	-0.01	-0.01	-0.01	-0.01	0.01	0.01
β_2	0.02	0.02	-0.01	-0.01	0.01	0.01

Table 4. 18 RMSE summaries for data generated by NB with low-dispersed datasets

Estimated parameters	N=100		N=500		N=1000	
	NB	GW	NB	GW	NB	GW
β_0	0.09	0.09	0.05	0.05	0.03	0.03
β_1	0.08	0.08	0.05	0.05	0.03	0.03
β_2	0.10	0.10	0.05	0.05	0.03	0.03

In conclusion, it can be seen from above analysis that the sample size also has a significant effect on the estimation process for datasets generated by NB distribution although not as large as on datasets generated by GW distribution. It can be seen from above tables that the values of estimated parameter are closer to the true value when sample size increases. The estimates of β_0, β_1 and β_2 always have a small bias and low standard errors for both models applied in datasets with relative large simple size.

The results of also show that the value of AIC and BIC for both models is almost the same and the bias and RMSE of estimated parameters are also very close, which means that the GW converges to the NB model. It can therefore be argued that the GW can approximate the data which were originally generated by a Negative Binomial distribution. Furthermore, the GW could technically be used over the NB when sample size is relatively large. However, the GW is more complex to estimate than the NB model.

In similar as Scenario 1, the plots in Figures 4.4-4.6 visualize the goodness-of-fit comparison between the generated (or observed) frequencies and the predicted frequencies from two different kinds of models applied in overdispersed datasets generated from NB distribution. It clearly shows that the goodness of fit is almost the same between the NB model and the GW regression model for these simulated datasets.

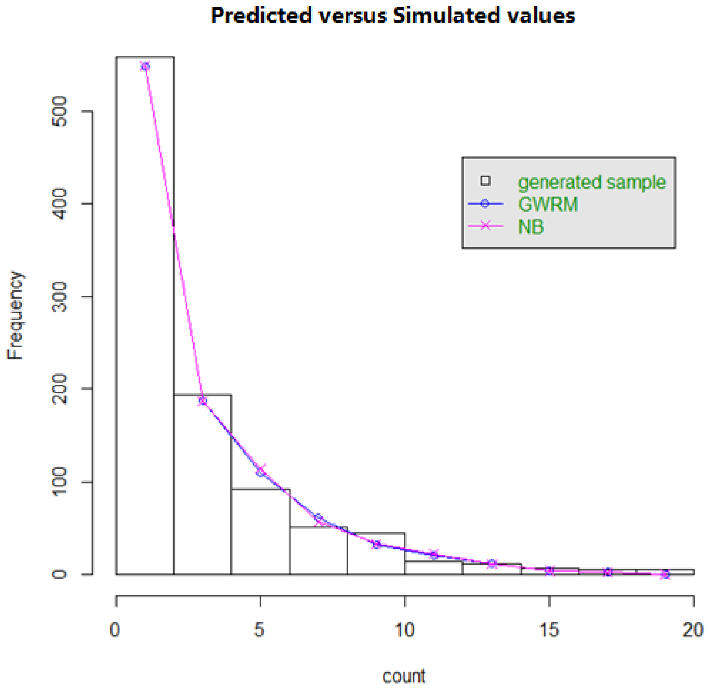

Figure 4. 4 Predicted versus simulated values for NB data Φ=2

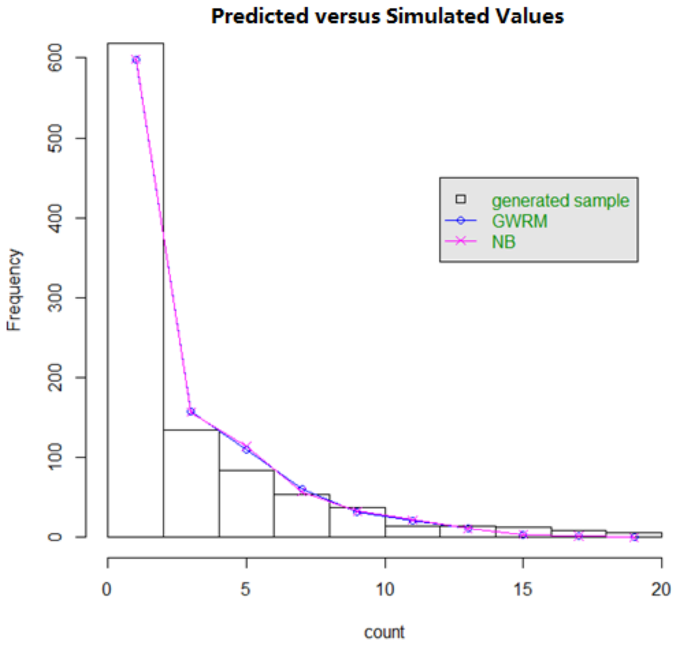

Figure 4.5 Predicted versus simulated values for NB data Φ=1

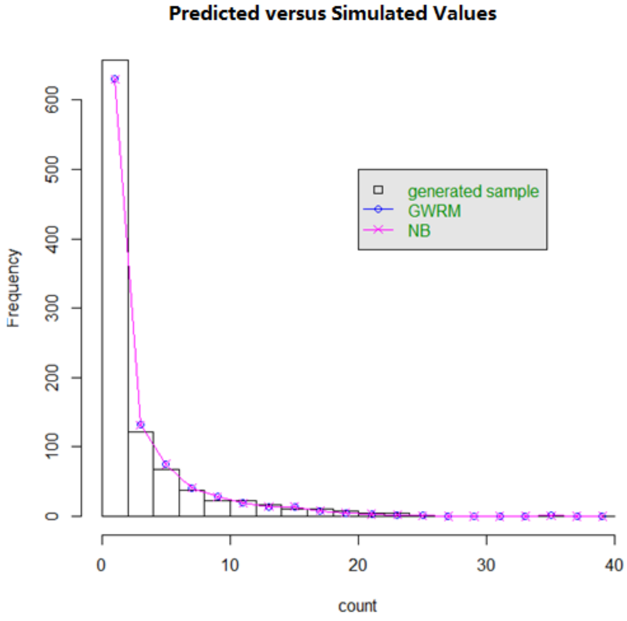

Figure 4.6 Predicted versus simulated values for NB data Φ=0.5

4.3 Scenario 3

The objective of this scenario consists of examining the performance of the GW and NB models when the simulated dataset is produced from a two-component finite Poisson mixture model.

4.3.1 Data Generation Method

For this scenario, the random variables (y) were simulated in the following steps:

Step 1: Set sample size, β and the parameters of two-mixture Poisson distribution to the required values.

Based on the two components' means, the FMP-2 random variable for site i was generated by introducing a mixing weight w. Thus, the random variable for the site i was generated from the Poisson ($\mu_{i,1}$) distribution with probability w and generated from the Poisson ($\mu_{i,2}$) distribution with probability 1- w. The datasets were generated by selecting the following values for these parameters: w =0.2 and β_0 =2, β_1 = 0.5, β_2 =-0.5 for Poisson ($\mu_{i,1}$) distribution to represent higher crash mean population and β_0 =0, β_1 = 0.5, β_2 =-0.5 for Poisson ($\mu_{i,2}$) distribution to represent lower crash mean population. The generated data could be classified as being highly dispersed and resemble empirical crash frequency plots which are commonly to be encountered in highway crash analysis. Sample size 100,500 and 1000 were analyzed separately for this simulation.

Step 2: Generate two covariates (X_{1i} , X_{2i}) from the standard normal distribution.

Step 3: Calculate the parameter by using assumed log-linear function.

Step 4: Generate the count variable y_i from the two-mixture Poisson distribution.

Step 5: The simulated datasets were then fitted by using FMP-2, GW and NB models.

Step 6: Repeat Steps 1 through 5 50 times and compare the average value of simulated results with the theoretical values and assess the general performance of the models.

4.3.2 Modeling Results

The simulation result of parameter estimation was shown in the following Tables 4.19-4.23. It can be seen in Table 4.19 that the values of estimated parameter are closer to the true value when sample size increases. The estimates of β_1 and β_2 have a smaller bias and lower standard errors when sample size above 500.

It also can be seen from the above table that GW performs better than the NB model for data generated by the two-component finite mixture of Poisson distribution, shown by the AIC and BIC values and the smaller confidence intervals. It could therefore be argued that the GW is more flexible than the NB for over dispersed crash data drawn from two heterogeneous populations.

Table 4.19 Parameter estimation results for both models using data generated by FMP-2 model

	N=100				N=500				N=1000			
	FMP-2		**NB**	**GW**	**FMP-2**		**NB**	**GW**	**FMP-2**		**NB**	**GW**
	Comp1	**Comp2**			**Comp1**	**Comp2**			**Comp1**	**Comp2**		
β_0	2.06 (0.10)	0.04 (0.12)	0.86 (0.12)	0.83 (0.12)	1.98 (0.05)	0.02 (0.06)	0.84 (0.04)	0.81 (0.06)	1.99 (0.03)	-0.01 (0.03)	0.82 (0.03)	0.81 (0.02)
β_1	-0.52 (0.11)	-0.55 (0.12)	-0.38 (0.12)	-0.42 (0.11)	-0.51 (0.05)	-0.51 (0.05)	-0.52 (0.05)	-0.51 (0.04)	-0.51 (0.03)	-0.51 (0.03)	-0.51 (0.02)	-0.49 (0.03)
β_2	0.44 (0.12)	0.47 (0.11)	0.61 (0.13)	0.44 (0.11)	0.48 (0.04)	0.49 (0.05)	0.48 (0.05)	0.48 (0.04)	0.51 (0.02)	0.51 (0.03)	0.49 (0.03)	0.49 (0.03)
w	0.18	0.82	-	-	0.19	0.81	-	-	0.2	0.8	-	-
ρ	-	-	-	13.1	-	-	-	6.53	-	-		5.27
k	-	-	-	1.82	-	-	-	2.01	-	-		2.23
Φ	-	-	1.08 (0.21)	-	-	-	0.97 (0.06)	-	-	-	1.01 (0.04)	-
-2LL	358 (8.26)		417 (8.68)	**382** (8.49)	1889 (38.26)		2141 (39.96)	**2090** (38.97)	3916 (82.13)		4104 (84.63)	**4052** (83.21)
AIC	372 (8.26)		425 (8.68)	**392** (8.49)	1903 (38.26)		2149 (39.96)	**2100** (38.97)	3930 (82.13)		4112 (84.63)	**4062** (83.21)
BIC	390 (8.26)		426 (8.68)	**406** (8.49)	1932 (38.26)		2153 (39.96)	**2121** (38.97)	3965 (82.13)		4128 (84.63)	**4086** (83.21)

NOTE: ( ) indicate the standard error of the estimate

Table 4.20 Bias summaries using data generated by FMP-2 models (component 1)

Estimated parameters	**N=100**			**N=500**			**N=1000**		
	FMP-2	**NB**	**GW**	**FMP-2**	**NB**	**GW**	**FMP-2**	**NB**	**GW**
β_0	0.06	-	-	-0.02	-	-	-0.03	-	-
β_1	-0.02	0.12	0.08	-0.01	-0.02	-0.01	-0.01	-0.01	0.01
β_2	-0.06	0.11	-0.06	-0.02	-0.02	-0.02	0.01	-0.01	-0.01

Table 4.21 RMSE summaries using data generated by FMP-2 models (component 1)

Estimated parameters	N=100			N=500			N=1000		
	FMP-2	NB	GW	FMP-2	NB	GW	FMP-2	NB	GW
β_0	0.12	-	-	0.05	-	-	0.04	-	-
β_1	0.11	0.17	0.14	0.05	0.05	0.05	0.03	0.02	0.03
β_2	0.13	0.17	0.14	0.05	0.05	0.05	0.03	0.03	0.03

Table 4.22 Bias summaries using data generated by FMP-2 models (component 2)

Estimated parameters	N=100			N=500			N=1000		
	FMP-2	NB	GW	FMP-2	NB	GW	FMP-2	NB	GW
β_0	0.04			0.02			-0.007		
β_1	-0.05	0.12	0.08	-0.01	-0.02	-0.01	-0.01	-0.01	0.01
β_2	-0.03	0.11	-0.06	-0.01	-0.02	-0.02	-0.01	-0.01	-0.01

Table 4.23 RMSE summaries using data generated by FMP-2 models (component 2)

Estimated parameters	N=100			N=500			N=1000		
	FMP-2	NB	GW	FMP-2	NB	GW	FMP-2	NB	GW
β_0	0.13	-	-	0.06	-	-	0.03	0.00	0.00
β_1	0.13	0.26	0.22	0.05	0.04	0.01	0.03	0.02	0.01
β_2	0.11	0.20	0.11	0.05	0.06	0.04	0.03	0.02	0.01

The goodness of fit was compared and presented in the following Figure 4.7: it clearly shows that the GW model provides better result than the NB model for overdispersed data drawn from two distinct Poisson distribution.

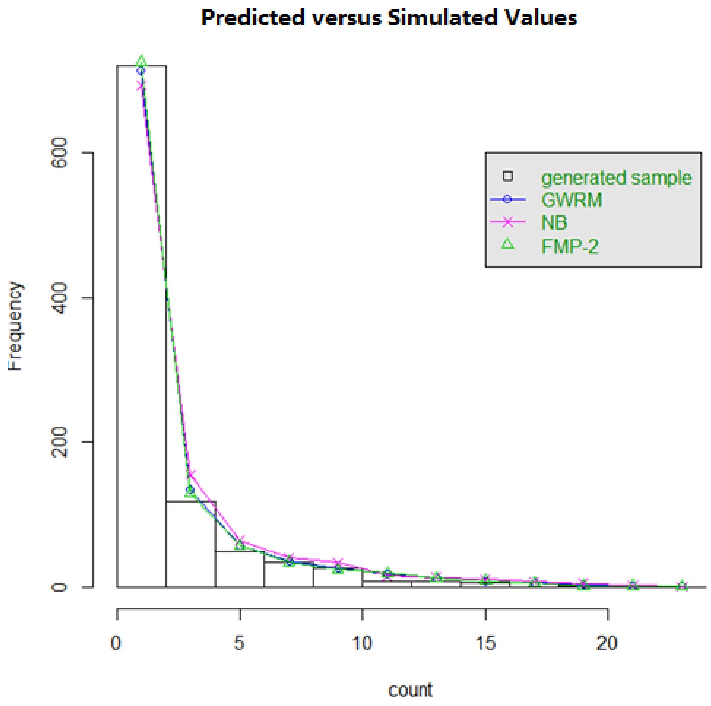

Figure 4.7 Predicted versus simulated values for FMP-2 data (N=1000)

4.4 Scenario 4

It is shown in the above section that the GW model provides better goodness of fit than the NB model for overdispersed data drawn from two distinct Poisson distribution. In this section, the simulated dataset is produced from a two-component finite Negative Binomial distribution for the purpose of comparing the performance of these two models on overdispersed data in further.

4.4.1 Data Generation Method

For this scenario, the random variables (y) were simulated in the following steps:

Step 1: Similar as the simulated data generated by two-mixture Poisson distribution, set sample size and the parameters of two-mixture NB distribution to the required values.

The datasets were generated by selecting the following values for these parameters: w =0.2 and β_0=2, β_1= 0.5, β_2=-0.5 $\phi_1 = 5$ for NB($\mu_{i,1}$) distribution to represent higher crash mean population and β_0=0, β_1= 0.5, β_2=-0.5 $\phi_2 = 5$ for NB($\mu_{i,2}$) distribution to represent lower crash mean population. Sample size 100,500 and 1000 were analyzed separately for this simulation.

Step 2: Generate two covariates (X_{1i} , X_{2i}) from the standard normal distribution.

Step 3: Calculate the parameter by using assumed log-linear function.

Step 4: Generate the count variable y_i from the two-mixture NB distribution.

Step 5: The simulated datasets were then fitted by using FMNB-2, GW and NB models.

Step 6: Repeat Steps 1 through 5 50 times and compare the average value of simulated results with the theoretical values and assess the general performance of the models.

4.4.2 Modeling Results

The simulation result of parameter estimation was shown in the following Tables 4.24-4.27. Similar as the results from above section, it can be seen in Table 4.24 that the values of estimated parameter are closer to the true value when sample size increases.

It also can be seen from the above table that GW performs better than the NB model for data generated by the two-component finite mixture of NB distribution because of the larger log likelihood values and smaller AIC and BIC values. It could therefore be further confirmed that the GW is more flexible than the NB for over dispersed crash data drawn from two heterogeneous populations.

Table 4.24 Parameter estimation results for both models using data generated by FMNB-2 model

	N=100				N=500				N=1000			
	FMP-2		**NB**	**GW**	**FMP-2**		**NB**	**GW**	**FMP-2**		**NB**	**GW**
	Comp 1	**Comp 2**			**Comp 1**	**Comp 2**			**Comp 1**	**Comp 2**		
β_0	1.76 (0.34)	-0.12 (0.13)	1.03 (0.14)	1.11 (0.19)	1.89 (0.13)	0.05 (0.04)	0.90 (0.06)	0.94 (0.08)	1.97 (0.08)	-0.03 (0.05)	0.82 (0.03)	0.89 (0.05)
β_1	-0.54 (0.23)	-0.55 (0.13)	-0.61 (0.12)	-0.57 (0.12)	-0.53 (0.12)	-0.47 (0.06)	-0.40 (0.06)	-0.46 (0.07)	-0.53 (0.04)	-0.53 (0.04)	-0.52 (0.02)	-0.49 (0.04)
β_2	0.42 (0.14)	0.46 (0.13)	0.26 (0.14)	0.36 (0.13)	0.46 (0.13)	0.46 (0.05)	0.42 (0.06)	0.45 (0.07)	0.52 (0.04)	0.52 (0.04)	0.48 (0.03)	0.49 (0.04)
w	0.17	0.83	-	-	0.18	0.82	-	-	0.20	0.80	-	-
ρ	-	-	-	4.52	-	-	-	3.22	-	-		3.31
k	-	-	-	2.81	-	-	-	2.46	-	-		2.27
Φ	3.89 (1.46)	4.36 (1.43)	0.74 (0.14)	-	4.45 (1.16)	4.35 (1.02)	0.71 (0.08)	-	4.68 (0.76)	4.78 (0.87)	1.01 (0.04)	-
-2LL	368 (9.01)		475 (10.25)	**426** (9.58)	1892 (48.31)		2115 (42.36)	**2081** (40.52)	3951 (83.24)		4160 (87.98)	**4124** (86.54)
AIC	386 (9.01)		483 (10.25)	**436** (9.58)	1910 (48.31)		2123 (42.36)	**2091** (40.52)	3969 (83.24)		4168 (87.98)	**4134** (86.54)
BIC	409 (9.01)		484 (10.25)	**450** (9.58)	1948 (48.31)		2127 (42.36)	**2112** (40.52)	4013 (83.24)		4184 (87.98)	**4160** (86.54)

Table 4.25 Bias summaries using data generated by FMNB-2 models (component 1)

Estimated parameters	N=100			N=500			N=1000		
	FMNB-2	NB	GW	FMNB-2	NB	GW	FMNB-2	NB	GW
β_0	-0.24	-	-	-0.02	-	-	-0.03	-	-
β_1	-0.04	-0.11	-0.07	-0.03	0.10	0.04	-0.03	-0.02	0.01
β_2	-0.08	-0.24	-0.14	-0.04	-0.08	-0.05	0.02	-0.02	-0.01

Table 4.26 RMSE summaries using data generated by FMNB-2 models (component 1)

Estimated parameters	N=100			N=500			N=1000		
	FMNB-2	NB	GW	FMNB-2	NB	GW	FMNB-2	NB	GW
β_0	0.42	-	-	0.13	-	-	0.09	-	-
β_1	0.23	0.16	0.14	0.12	0.12	0.08	0.05	0.03	0.04
β_2	0.16	0.28	0.19	0.14	0.10	0.09	0.04	0.04	0.04

Table 4.27 Bias summaries using data generated by FMNB-2 models (component 2)

Estimated parameters	N=100			N=500			N=1000		
	FMNB-2	NB	GW	FMNB-2	NB	GW	FMNB-2	NB	GW
β_0	-0.12	-	-	0.05	-	-	-0.007	-	-
β_1	-0.05	-0.11	-0.07	0.03	0.1	0.04	-0.03	-0.02	0.01
β_2	-0.04	-0.24	-0.14	-0.04	-0.08	-0.05	0.02	-0.02	-0.01

Table 4.28 RMSE summaries using data generated by FMNB-2 models (component 2)

Estimated parameters	N=100			N=500			N=1000		
	FMNB-2	NB	GW	FMNB-2	NB	GW	FMNB-2	NB	GW
β_0	0.18	-	-	0.06	-	-	0.05	-	-
β_1	0.14	0.16	0.14	0.07	0.12	0.08	0.05	0.03	0.04
β_2	0.14	0.28	0.19	0.06	0.10	0.09	0.04	0.04	0.04

The goodness of fit was compared and presented in the following Figure 4.8: it clearly shows that the GW model provides better result than the NB model for overdispersed data composed of two NB distributions.

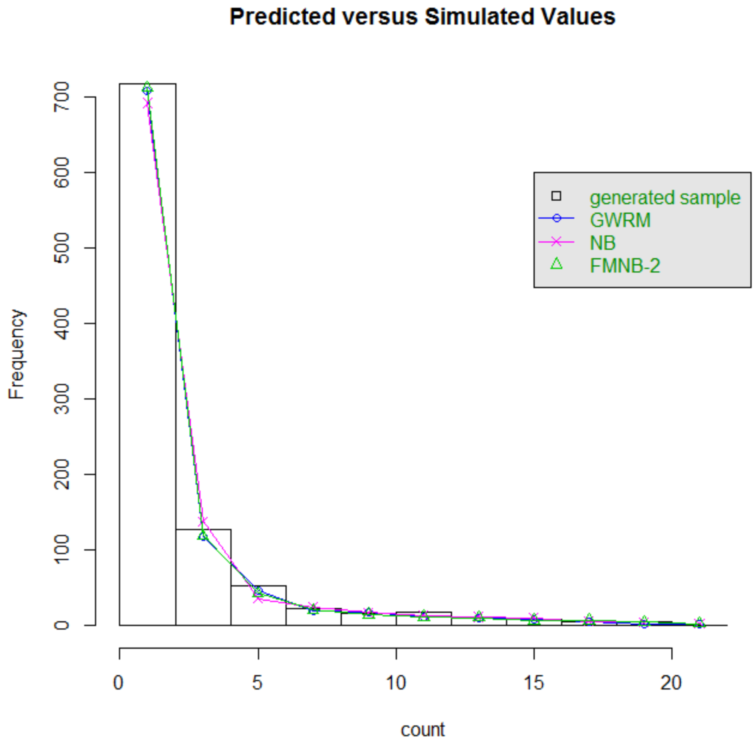

Figure 4.8 Predicted versus simulated values for FMNB-2 data (N=1000)

4.5 Chapter Summary

This chapter has described the compared results of performance of both GW and NB models on different kinds of datasets. The simulation results show several key findings. The first simulated scenario illustrates the good numerical approximation of GW regression models can provide and the poor capability of the traditional NB model when the underlying distribution comes from GW distribution. In addition, it is shown that the sample size and sample mean have a significant effect on the estimation process. The parameter accuracy will increase as the sample size and sample mean increase, as expected.

Secondly, simulated datasets were used to illustrate the appropriateness of the GW model specifications when the data were actually generated from a Negative Binomial distribution with different levels of dispersion.

Finally, to examine potential bias with using either NB distribution or GW distribution to compare performance, simulated datasets were also generated using a two-components finite mixture of Poisson distribution and two-components finite mixture of NB distribution . It was shown that GW performed better than NB model for datasets generated from two heterogeneous populations. The next chapter focuses on discussion of performance of GW models on empirical crash datasets and investigation of source of overdispersion for these datasets.

CHAPTER V

EMPIRICAL CRASH DATA ANALYSIS

Simulated scenarios were analyzed in the previous chapter to illustrate the good performance of the GW model in accounting for over-dispersed data. It is necessary to examine the underlying assumption in the GW model in further when this model is applied to empirical crash data. It is assumed that all road segments or intersections are not exposed to exactly the same external risk of accident. Differences in exposure to external factors from one segment or intersection to another are known as differences in accident liability, as distinguished from constitutional or internal differences which are known as differences in proneness. In practice, the effects of proneness and liability are confounded and inseparable when the Negative Binomial is fitted. However, these two parts of variance can be identified by using the GW model.

The objective of this chapter is to apply the GW model to actual vehicle crash data and to demonstrate the model's ability to discern the sources of variance in the data. The results of these models are compared to those produced from the standard NB regression model in terms of both GOF and information about sources of variance.

Two kinds of datasets are considered for this application: intersection crash data (Section 5.1) and segment crash data (Section 5.2). For intersection crash data, both the GW and NB models were used to analyze the signalized intersection crash data obtained from Toronto, Ontario.

For segment crash data, this research utilized several multilane segments of crash data for highways in Texas, Indiana, and Michigan; all of these segments have been analyzed by other researchers for other studies (Geedipally and Lord, 2011). Compared to the intersection dataset contains only traffic flow variables, the segment dataset has more covariates (such as median width and shoulder width) as well as traffic flow. Therefore, it is possible to examine whether the performance of the GW would work better than the NB both in a model with few covariates and in a more fully-specified model. It should be noted that the Michigan crash dataset was utilized to compare the performance of the NB, NB-L, and GW regression models to improve the completeness of this research.

5.1 Intersection Crash Data Analysis

This section presents the dataset and analysis results for the intersection crash data.

5.1.1 Toronto Data Description

In order to test the performance of the GW model for real intersection crash data, data collected at urban 4-legged signalized intersections in Toronto were used. Even though the dataset was collected a long time ago, there are two main advantages to using this dataset. First, this dataset has been analyzed by many traffic safety researchers for different purposes and the quality of this dataset has been confirmed (Lord, 2009; Persaud et al., 2002; Miaou and Lord, 2003; Lord et al., 2008). Secondly, the dataset contains only two covariates: traffic flows for major and minor approaches. Many factors that may affect the number of crashes have not been observed or included in the dataset. Therefore, it is meaningful to apply the GW model to this dataset to examine the different sources of variance of crashes by dividing the total variances into three parts: randomness, proneness, and liability. Such an application is one of the major strengths of the GW model.

The summary statistics for the dataset are listed in Table 5.1. There are 868 intersections in total; 10,030 reported crashes occurred on the road network. The number of crashes on each site varies from 0 to 54, and the sample mean and sample variance equals 11.56 (crashes/intersection) and 100 (crashes/intersection) separately.

Since the type of crashes includes both intersection-related and non-injury crashes, the value of the sample mean is high. The observed crash frequency plot is shown in Figure 5.1. The traffic volumes vary from 5,469 to 72,178 vehicles/day for major approaches, and from 53 to 42,644 vehicles/day for minor approaches. The more detailed descriptions of the dataset are described in Lord (2000).

Table 5.1 Summary statistics for intersection dataset

Variable	Maximum	Minimum	Average	Standard Deviation
Major Approach (F1)	72718	5469	28045	10660
Minor Approach(F2)	42644	53	11010	8599
Crashes	54	0	11.56	10.02

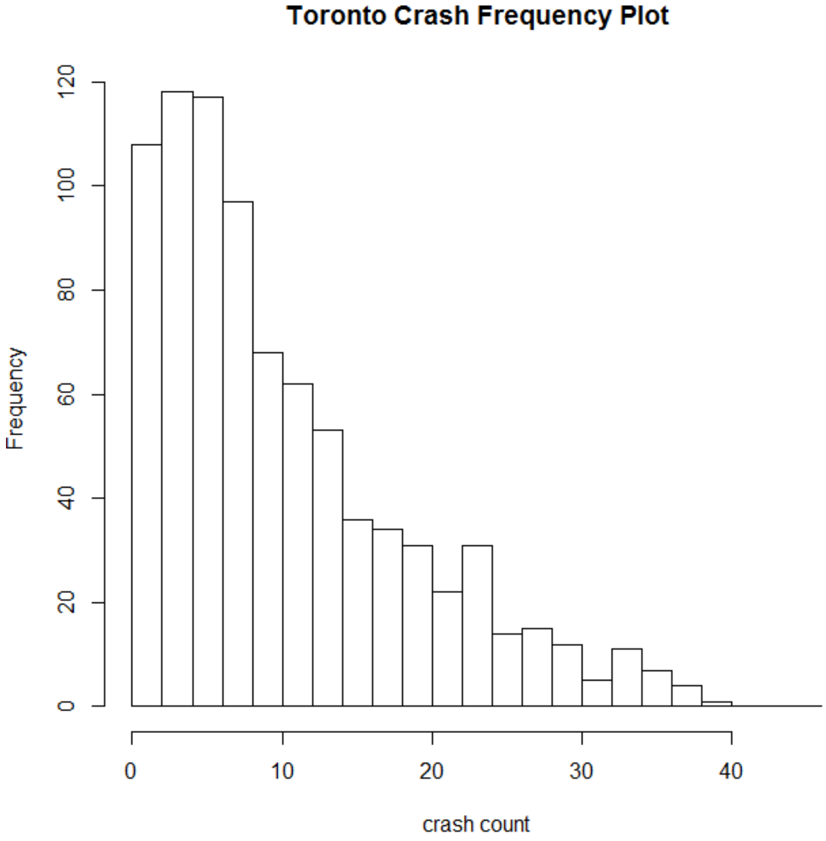

Figure 5. 1 Plot of Toronto crash frequency

5.1.2 Mean Functional Form Discussion

Although there are a lot of factors that may affect the number of crashes near intersections, some transportation safety researchers still prefer models with the only traffic flow variable over models including many other covariates because they can be estimated and calibrated easily (Persaud et al., 2002; Lord and Bonneson, 2005). However, this kind of model may sometimes be significantly affected by the bias generated from omitted variables (Lord and Mannering, 2010). Miaou and Lord (2003) compiled a list of five commonly used functional forms based on previous studies:

$$\begin{aligned}
\mu_i &= \beta_0 (F_{1i} + F_{2i})^{\beta_1} \\
\mu_i &= \beta_0 F_{1i}^{\beta_1} F_{2i}^{\beta_2} \\
\mu_i &= \beta_0 (F_{1i} F_{2i})^{\beta_1} \\
\mu_i &= \beta_0 (F_{1i} + F_{2i})^{\beta_1} (F_{2i} / F_{1i})^{\beta_2} \\
\mu_i &= \beta_0 F_{1i}^{\beta_1} F_{2i}^{\beta_2} \exp(\beta_3 F_{2i})
\end{aligned} \tag{5.1}$$

Among these functional forms, the following functional form is the most popular, and has been most favored by transportation safety modelers for use in modeling crash data at intersections. It should be noted that it does not appropriately fit the data near the boundary conditions since the crash mean should not be zero unless both F1 and F$_2$ are zero.

$$\mu_i = \beta_0 F_{1i}^{\beta_1} F_{2i}^{\beta_2} \tag{5.2}$$

To overcome the boundary value limitation, Miaou and Lord (2003) proposed an alternative form (see Equation 5.3) which represents two different risk levels for vehicles entering the two approaches.

$$\mu_i = F_{1i} \exp(\beta_0 + \beta_1 F_{2i}) + F_{2i} \exp(\beta_0 + \beta_2 F_{1i}) \tag{5.3}$$

In this chapter, the empirical data was fitted based on the most popular functional form shown in Equation (5.2).

5.1.3 Modeling Results

The analysis of the modeling results is divided into three parts: a goodness of fit analysis, a source of variance analysis, and a covariate sensitivity analysis.

5.1.3.1 Goodness of Fit Analysis

The goodness of fit results of the NB and GW models are presented in the following table. The coefficient values of both covariates in both of the models have the same sign and are close to each other. The crashes increase with the increase in traffic flow in both models. Both models indicate that the crash risk increases at a decreasing rate as traffic flow increases because the coefficient is below 1.

Table 5.2 Modeling results for the Toronto data

Variable	**NB**	**GW**
Intercept	-10.24(0.65)	-10.24(0.65)
Ln(F1)	0.62(0.06)	0.62(0.01)
Ln(F2)	0.68(0.03)	0.69(0.02)
ρ	-	8.2e+05
K	-	7.15
$\delta = 1/\phi$	0.14(0.01)	-
-2LL (the smaller the better)	5069	5069.20
AIC(the smaller the better)	5077	5079.20
BIC(the smaller the better)	**5096**	5103.10
MAD	**4.14**	4.18
MSPE	33.50	**32.87**
Pearson χ^2	2343.70	**2127.30**

NOTE: ( ) indicates the standard deviation

$$MAD = \frac{1}{n} \sum_{i=1}^{n} | y_i - y_i |$$

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$\chi^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{y_i}$$

The goodness of fit for both models can also be compared in the following figure by comparing the observed and predicted frequencies of each crash count outcome using the same method mentioned in Chapter IV.

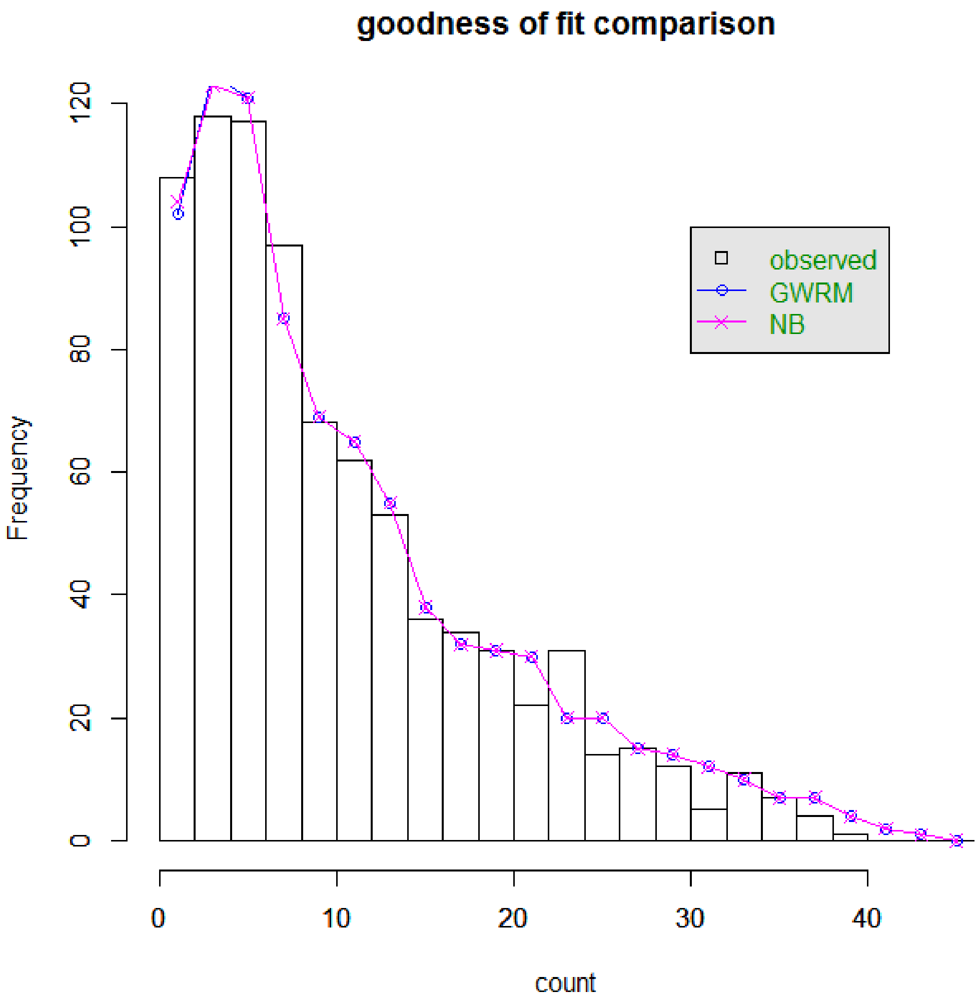

Figure 5. 2 Goodness of fit comparison between NB and GW

Based on the modeling results shown in Table 5.2, almost all goodness of fit statistical criteria are close to each other for both kinds of models; this indicates that both models produce a satisfactory fit for this intersection dataset. In detail, the values of AIC, BIC and MAD for the NB model are slightly smaller than the values for the GW model, and the values of MSPE and Pearson χ^2 for the NB model are slightly larger than those of the GW model.

It is also shown in Figure 5.2 that the difference between the NB and GW models is very small, and the discrepancy of predicted frequencies between the two models is almost negligible. It should be noted that although the standard NB model works very well for this intersection dataset, it does not mean that the NB model will produce the same satisfactory fit as the GW model for the other crash datasets.

5.1.3.2 Sources of Variance Analysis

As mentioned in previous sections, although the standard NB model works very well, it cannot provide adequate information about the potential sources of over-dispersion. In this case, the GW model can be used to examine the possible existence of different sources of variance for crashes in each intersection and to identify the quantity of each part of variance.

From a statistical viewpoint, the number of crashes occurring in each intersection is a discrete count variable with a tendency towards over-dispersion. This excess variability is different from the Poisson model which considers only the effect of randomness. The variance also is caused by external factors observable by covariates that significantly influence the risk of crashes (for instance, in this intersection dataset, the ADT on each approach of intersections). These covariates would determine what in accident theory is called liability.

When it comes to crashes that occur in each intersection, there is a feature that is not related to external factors but instead is associated with the internal characteristics of each intersection that are difficult to observe (that is, with their internal probability to cause accident, which in accident theory is called the proneness of each intersection). There are several internal characteristics of intersections related to that intersection's proneness such as the friction of the road surface, the damage condition of intersection, the maintenance condition of the traffic signs and the pavement markings at the intersection, the operation of traffic signals at the intersection, and so on.

In this context, it can be seen that the proposed GW model is capable of distinguishing these three sources of variability and of providing more information about the data than any other regression model (such as the Poisson or the Negative Binomial models) (Rodríguez et al., 2009).

Specific to this crash dataset, the GW model assumes that proneness represents over-dispersion due to between-intersection variations in their internal probability of causing accidents with the same ADT, while liability represents over-dispersion due to missing covariates which would affect intersections with the same ADT.

Table 5.3 shows the variance of each component for this crash dataset. Each variance is quantified according to Table 3.1 and the GW modeling results shown in Table 5.2.

Table 5. 3 Variance of each component for the Toronto crash data

Source of variability	Variance	Variance rate
Randomness	μ_x	$\frac{7.15}{7.15+\mu_x}$
Liability	0	0
Proneness	$\frac{{\mu_x}^2}{7.15}$	$\frac{\mu_x}{7.15+\mu_x}$
Total	$(\mu_x+\frac{\mu_x^2}{7.15})$	1

Figure 5.3 shows the fractions of variance attributed to randomness, liability, and proneness for different means of crashes μ_x which were determined by all of the values of the covariates included in the model. The relationships among the covariate ADT and the fractions of each component are also indicated in Figures 5.4 and 5.5.

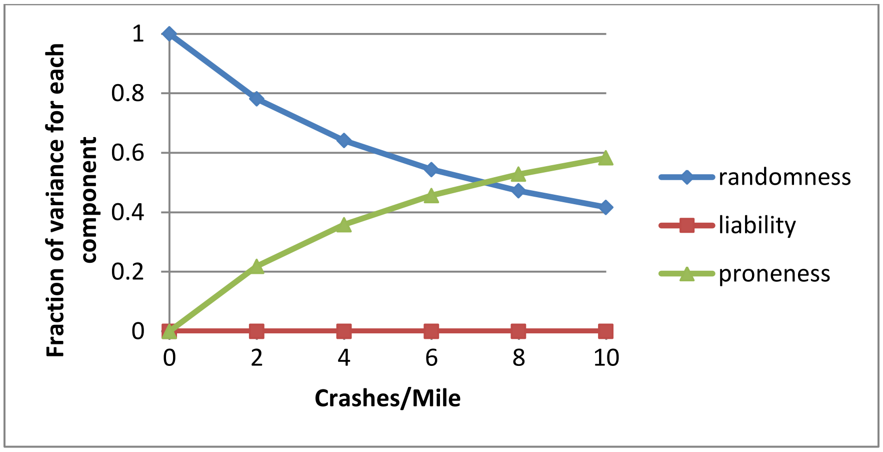

Figure 5. 3 Relationships among the fractions of variance for each component and mean of each crash

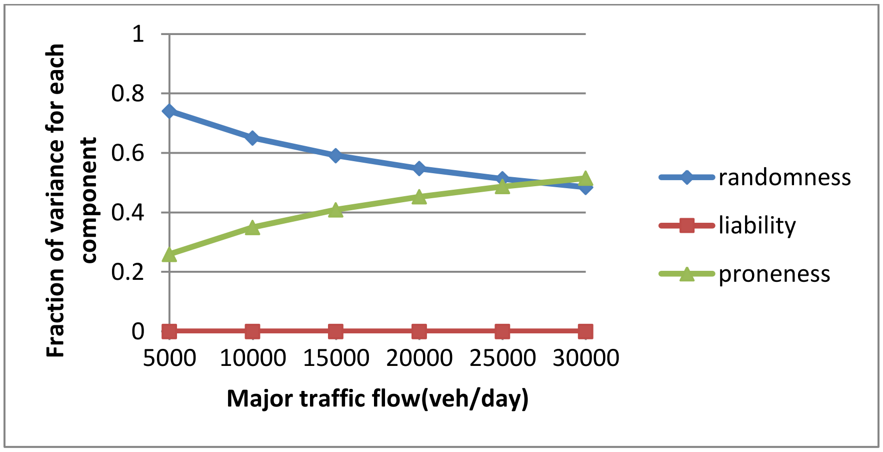

Figure 5.4 Relationships between fractions of variance for each component and major traffic flow (F2=5000)

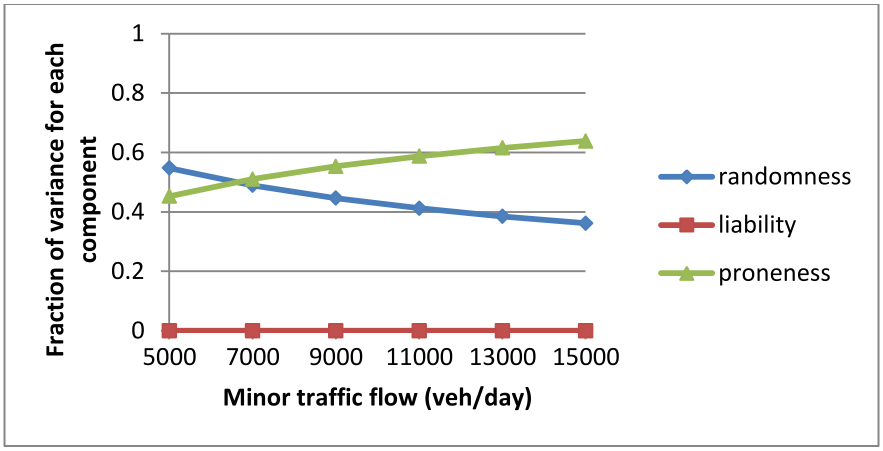

Figure 5. 5 Relationships between fractions of variance for each component and minor traffic flow (F1=20000)

It can be seen in Figures 5.3 to 5.5, the randomness decreases from 1 to 0.41 as the mean of crashes in each intersection increases from 0 to 10, which means that the randomness decreases when the ADT increases as the proneness increases. The fraction of proneness is more important in intersections with a higher ADT, which means that the source of over-dispersion comes more from proneness for those intersections with a higher ADT. In the major approach, the fraction of proneness increases from 0.25 to 0.51 with an increase in the average daily traffic volume from 5,000 to 30,000; in the minor approach, the fraction increases from 0.45 to 0.63 with an increase in the average daily traffic volume from 5,000 to 15,000. As was mentioned earlier, several internal characteristics of intersections, such as the friction of the road surface, the damage condition of the intersection, the maintenance condition of the traffic signs and pavement markings, and the operation of traffic signals can be related to the proneness. Therefore, the effects of these internal characteristics on the over-dispersion of traffic crashes seem to be more significant for the road segments with a higher ADT. Accordingly, for those segments with a lower mean of crashes and a lower ADT, the source of over-dispersion comes more from randomness. It should be noted that the liability in this dataset is always equal to 0 because of the extremely high value of ρ (according to the modeling results). As has been mentioned before, liability was determined by a set of external factors observable by covariates in the model. The external factors included in this model are the ADTs in both approaches. The results indicate that the source of variance is not caused by the two external factors included in the GW model. Therefore, it would be useful for traffic engineers to take effective measures to minimize the variance of traffic crashes based on a better understanding of the source of the variance. Traffic engineers should pay more attention to factors such as the damage condition of intersections and the maintenance condition of traffic signs and pavement markings, especially in intersections with a higher ADT, which is the main source of variance of crashes when a dataset exhibits a highly over-dispersed characteristic.

5.1.3.3 Covariate Sensitivity Analysis

The functional form for modeling the Toronto data is illustrated in Equation 5.2. Therefore, the sensitivity analysis for the covariate ADT can be examined by using the following equation:

$$\frac{\partial \log(\mu)}{\partial ADT} = \beta_1 \frac{1}{ADT_i} \tag{5.4}$$

Figure 5.6 shows the sensitivity of covariates F1 and F2 for both the NB and GW models. It can be seen that these two models exhibit quite similar trends. When the ADT is below 20,000 veh/day, one unit increase in ADT would result in a significant increase in the logarithm form of the estimated crash. However, this increasing rate becomes smaller when the ADT becomes larger.

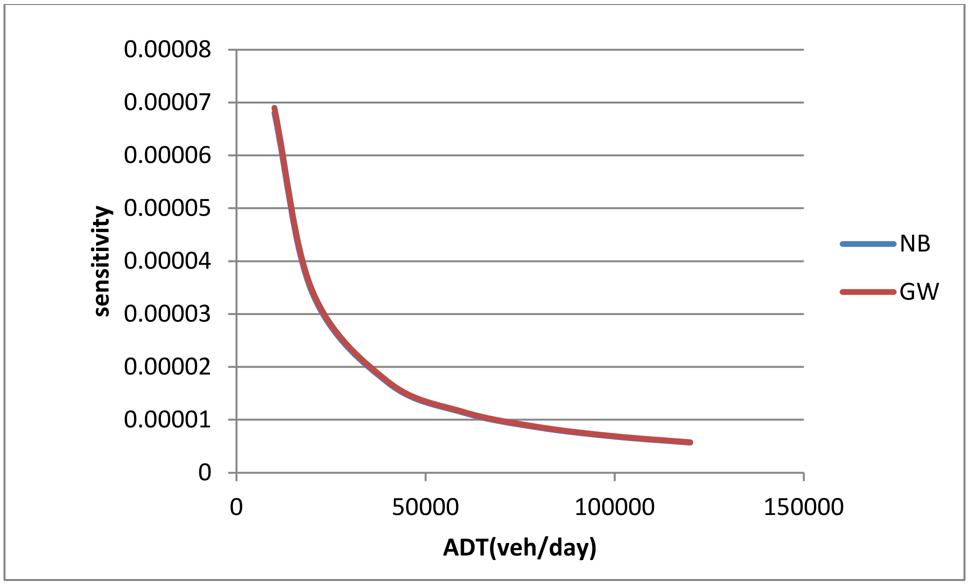

Figure 5. 6 Sensitivity analysis of covariate F1 for Toronto data

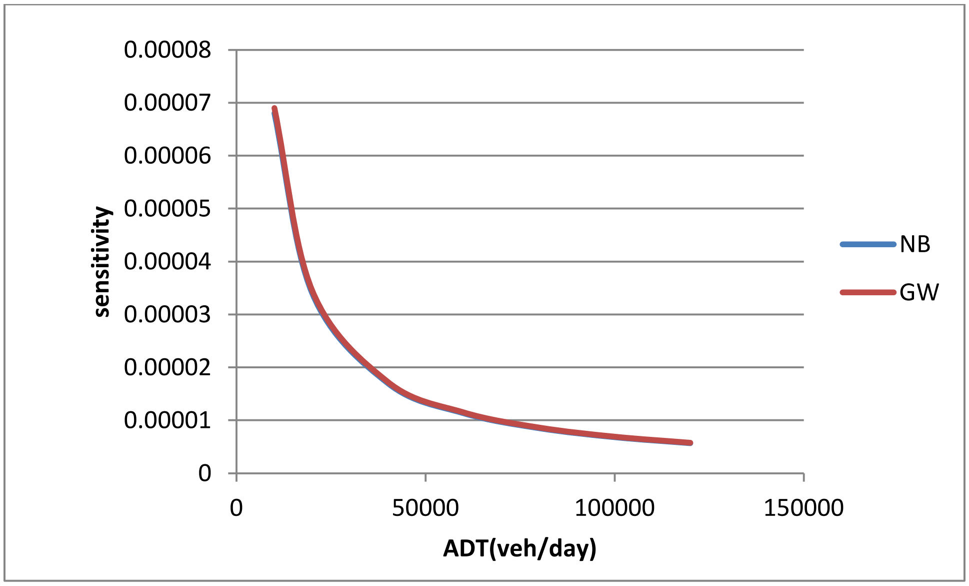

Figure 5. 7 Sensitivity analysis of covariate F2 for Toronto data

5.2 Road Segment Data Analysis

This section presents the dataset and the analysis results for the over-dispersed segment crash data.

5.2.1 Texas Crash Data Analysis

This dataset was collected from 4-lane, undivided rural segments in Texas. This dataset contains crash data collected from 1,499 undivided rural segments in Texas and was collected as a part of the NCHRP 17-29 research project (Lord et al., 2008). This dataset also has been analyzed by other researchers (see, e.g., Cheng et al., 2011). The length of segment ranges from 0.10 to 6.28 miles and the mean of the segment length is 0.55 mile. The mean and the variance of the crashes are 2.84 and 32.4, respectively. The summary statistics for the Texas data are indicated in Table 5.4.

Table 5.4 Summary statistics of the characteristics of the Texas data

Variable	**Minimum**	**Maximum**	**Mean(SD)**	**Sum**
Number of crashes	0	97	2.84(5.69)	4253
Average daily traffic	42	24800	6613(4010)	-
Lane width(feet)	9.75	16.50	12.57(1.59)	-
Total shoulder width (feet)	0	40	9.96(8.02)	-
Segment length(miles)	0.10	6.28	0.55(0.67)	830.49

5.2.1.1 Modeling Estimation

The following commonly used functional form for crash analysis was used in both models.

$$\mu_i = \beta_0 \times L_i \times F_i^{\beta_1} \times y \times e^{\sum_{j=2}^{n} X_{ij}\beta_j} \qquad (5.5)$$

Where,

μ_i = the estimated number of crashes per year for site i ;

F_i = vehicles per day (ADT) for segment i ;

L_i = length of segment i in miles;

y = number of years of crash data;

X_{ij} = a series of covariates (e.g., shoulder width, lateral clearance, etc.) for site i ;

n = number of covariates; and,

$\beta_1, \beta_2 .. \beta_n$ = estimated coefficients.

The data was then fitted based on these two functional forms with the GW model and the NB model to compare the performance of both models.

5.2.1.2 Goodness of Fit Analysis

The modeling results of the NB and GW models are presented in this section. Three covariates are considered in this analysis. They are: average daily traffic, lane width, and shoulder width. Segment length was considered an offset term in order to stay consistent with previous research (Geedipally and Lord, 2011). The coefficient values of all three covariates in both models have the same sign and are close to each other. In both models, the crashes decrease with an increase in lane width and shoulder width. The NB model shows that crashes increase almost linearly with an increase in traffic flow, while the GW model indicates that the crash risk increases at a slower rate as traffic flow becomes higher because the coefficient for the flow parameter is below 1. It can be seen from the goodness of fit statistics that the GW model performed much better than the NB model because the values of AIC, BIC, MAD, MSPE, and Pearson χ^2 obtained from the GW model are smaller than the values obtained from the NB model.

Table 5. 5 Modeling results for the Texas data

Variable	NB	GW
Intercept	-5.89(0.56)	-6.81(0.59)
Ln(ADT)	1.0063(0.06)	0.95(0.06)
LW	-0.1316(0.02)	-0.097(0.02)
SW	-0.0316(0.01)	-0.019(0.01)
ρ	-	3.01
K	-	1.71
Φ	0.66(0.03)	-
-2LL (the smaller the better)	6015.40	**5936.20**
AIC(the smaller the better)	6025.40	**5948.20**
BIC(the smaller the better)	6051.90	**5980.0**
MAD	2.75	**2.44**
MSPE	32.8	**28.40**
Pearson χ^2	12148.30	**11156.40**

NOTE: ( ) indicates the standard deviation

The goodness of fit for both models is also presented in Figure 5.8 by comparing the observed and predicted frequencies of each crash count outcome.

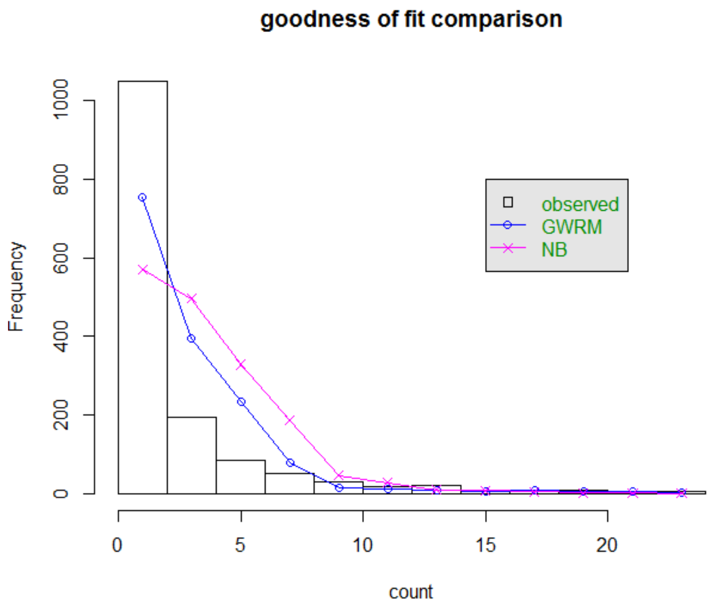

Figure 5.8 Goodness of fit comparison between the NB and GW models for Texas data

5.2.1.3 Sources of Variance Analysis

As mentioned above in the section discussing intersection crash data analysis, the other more important advantage of the GW model is that it can provide more information about the sources of variance for crashes occurring in each segment.

The liability was determined by a set of external factors observable by covariates including the ADT, the shoulder width, and the lane width in this dataset. When it comes to crashes occurring on each segment, there is a factor that is not related to external factors but is associated with the internal characteristics of each segment (characteristics which are difficult to observe). This is defined as the proneness of each segment. There are several internal characteristics of road segments that are related to the segments' proneness. For example, the friction of the road surface, the damage condition of the road segments, the maintenance condition of the rumble strips or pavement markings on the road segments, and so on.

Specific to this crash dataset, the GW model assumes that the proneness represents the over-dispersion due to between-segment variations in their internal probability to cause accidents with the same ADT, lane width, and shoulder width, while the liability represents the over-dispersion due to missing covariates which would affect identically those segments with the same ADT, lane width, and shoulder width.

Table 5.6 shows the variance of each component for this crash dataset according to Equation 3.1.

Table 5.6 Variance of each component for Texas crash data

Source of variability	Variance	Variance rate
Randomness	μ_x	$\frac{0.4642}{1.71+\mu_x}$
Liability	$2.683\mu_x$	$\frac{1.2457}{1.71+\mu_x}$
Proneness	$2.1539\mu_x{}^2$	$\frac{\mu_x}{1.71+\mu_x}$
Total	$3.683(\mu_x+\frac{\mu_x^2}{1.71})$	1

Figure 5.9 shows the fractions of variance attributed to randomness, liability, and proneness for different crash means μ_x which are determined by all the values of the covariates included in the model. Figure 5.10 illustrates the relationship between the covariate for the average daily traffic and the fractions of each component for a shoulder width equal to 16 feet (8 feet on each side) and the lane width equal to 12 feet. These two values represent the largest percentage of segments with these characteristics. It should be noted that the relationships between other covariates included in the model and the fractions of each component were also investigated and are presented in Appendix B. It can be seen in those figures in Appendix B that the sources of variance are highly dependent on traffic volumes compared to other variables included in the model.

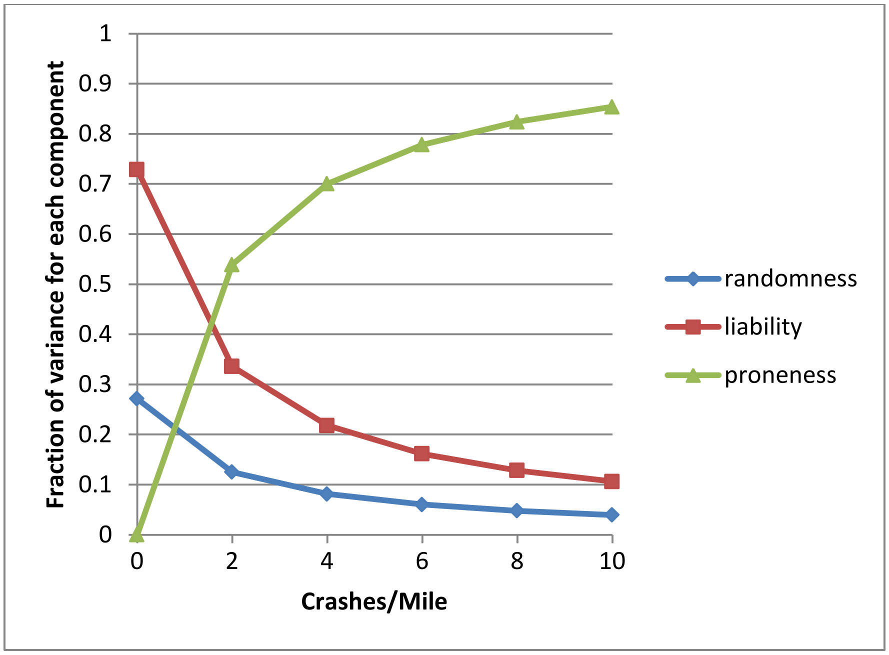

Figure 5.9 Relationships between the fractions of each component and the means of crashes for Texas data

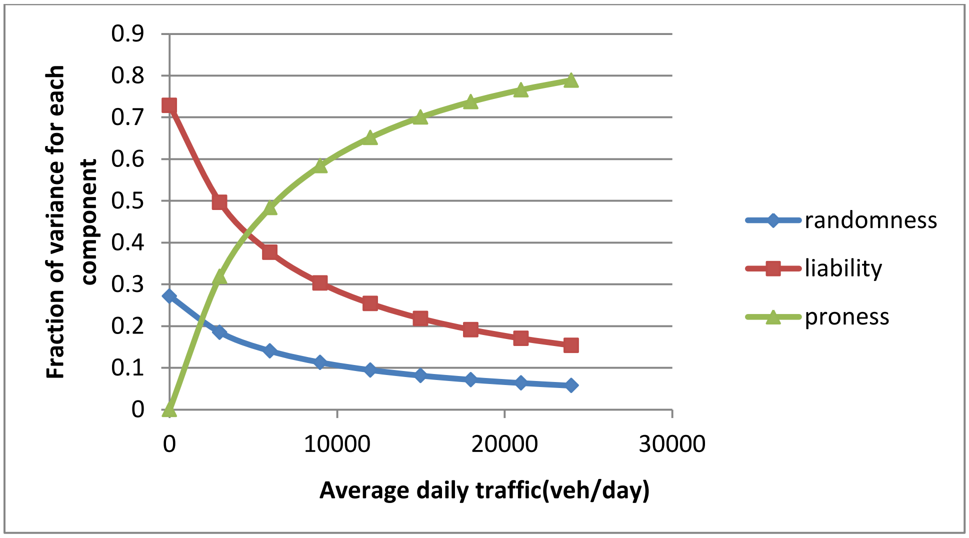

Figure 5.10 Relationships between the fractions of each component and the average daily traffic for Texas data

It can be seen from Figures 5.9 and 5.10 that randomness and liability decrease from 0.27 to 0.03 and from 0.72 to 0.1 as the mean of the crashes that occur on each segment μ_x increases from 0 to 10, which means that both randomness and liability decrease (whereas proneness increases) when the ADT increases and the shoulder width or lane width decreases. It is well-known that μ_x increases with an increase in the length of the segment, and randomness and liability decrease (whereas proneness increases) with the length of segment.

The fraction of proneness is more important for segments with a higher ADT, which means that the source of over-dispersion comes more from the proneness of those segments. The fraction of proneness increases from 0 to 0.79 with an increase in the ADT from 0 to 24,000. As mentioned above, several internal characteristics of the road segments such as the pavement friction, damage condition, maintenance condition of the rumble strips or pavement markings, and signs on the road segments can be related to the segment's proneness. Therefore, the effect of these internal characteristics on the over-dispersion of traffic crashes seems to be more significant on segments with a higher ADT. Accordingly, for those segments with a lower crash mean and a lower ADT, the source of over-dispersion comes more from the randomness and liability caused by the covariates included in the model.

5.2.1.4 Covariate Sensitivity Analysis

Figure 5.11 shows the sensitivity of covariates F1 and F2 for both the NB and GW models. It can be seen that those two models still have quite similar trends. When the ADT is below 15,000 veh/day, a one unit increase in ADT would result in a significant increase in the logarithm form of the estimated crash. However, this increasing rate becomes smaller when the ADT becomes larger.

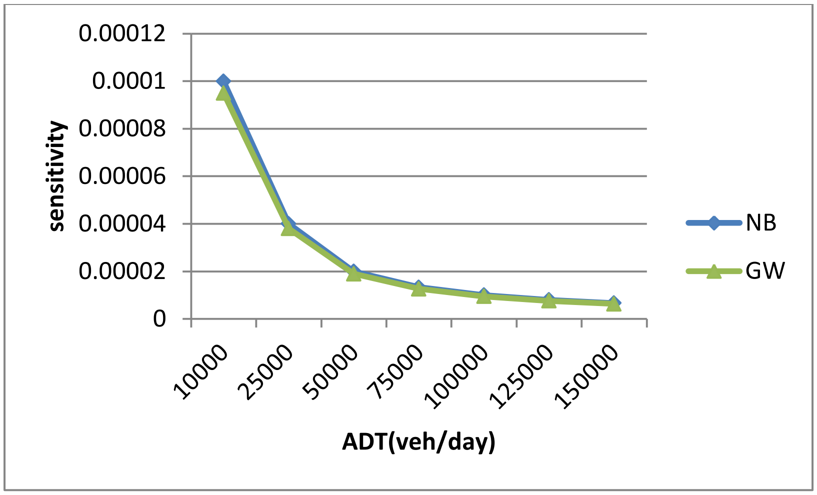

Figure 5.11 Sensitivity analysis of covariate ADT for the Texas data

5.2.2 Indiana Crash Data

This dataset was also collected from 1995 to 1999 and includes traffic and other covariates for 338 rural road segments in Indiana. The covariates included in this dataset are more than above those of the Texas dataset. The reason for using this dataset is similar to that of the Toronto intersection dataset. This Indiana dataset has been analyzed by other researchers and confirmed to be of good quality (Anastasopoulus et al., 2008; Washington et al., 2011). It should be noted that 120 of the 338 segments did not have any reported crashes over the five-year study period. For a more detailed list of variables, refer to Washington et al. (2011). There are more variables included in this dataset than in the above-mentioned Texas crash data.

Table 5.7 Summary Statistics for the Indiana Data

Variable	Min	Max	Average(std.dev)	Total
Number of crashes (5 years)	0	329	16.97(36.3)	5737
Average daily traffic over the 5 years	9442	143422	30237(28776.4)	-
Minimum friction reading in the road segment over the 5-year period (Friction)	15.9	48.2	30.51(6.67)	-
Pavement surface type	0	1	0.77(0.42)	-
Median width (in feet)	16	194.7	66.98(34.17)	-
Presence of median barrier (1 if present, 0 if absent) (BARRIER)	0	1	0.16(0.37)	-
Interior rumble strips (RUMBLE)	0	1	0.72(0.45)	-
Segment length (in miles)	0.009	11.53	0.89(1.48)	300.09

5.2.2.1 Goodness of Fit Analysis

Table 5.8 summarizes the parameter estimation results for the dataset. The segment length variable is still considered an offset during the estimation process. It can be seen from the Table 5.8 that the coefficients for the parameters of traffic flow are below 1 for both the GW and NB models, which indicates that the crash risk increases at a decreasing rate as the traffic flow increases. It is shown in Table 5.8 that the GW model performs better in terms of fit than the NB model. The estimated coefficients of all the covariates between the two models have the same sign which indicates similar results with regards to the effects of these variables, though there are some obvious differences between the values of these variables. It also can be seen from the results that the standard errors for the estimated coefficients are slightly larger for the GW model as compared to those from the NB model because the GW model is a multi-level hierarchical model that includes more parameters than a simple parametric model. As a result, the effective degrees of freedom could be smaller leading to increased standard errors (Geedipally and Lord, 2012). However, it is still beneficial to use the GW model to improve the predictive modeling ability because of its better ability to explain any over-dispersion.

Table 5.8 Modeling results for the Indiana data

Variable	**NB**	**GW**
Intercept	0.46(1.98)	-2.81(2.6)
Ln (ADT)	0.46(0.08)	0.70(0.14)
FR	-0.03(0.01)	-0.03(0.01)
PS	0.34(0.28)	0.44(0.34)
MW	-0.02(0.002)	-0.01(0.003)
MB	-3.75(0.38)	-6.24(0.74)
IRS	-0.13(0.28)	-0.02(0.33)
ρ	-	2.85
k	-	0.55
ϕ	2.71	-
-2LL (the smaller the better)	2116	**2086.3**
AIC	2132	**2104.3**
BIC	2162	**2138.7**
MAD	17.9	**17.3**
MSPE	332.8	**305.3**
Pearson χ^2	1517.5	**1470.6**

The goodness of fit of both models is also presented in Figure 5.12 by comparing the observed and predicted frequencies of each crash count outcome.

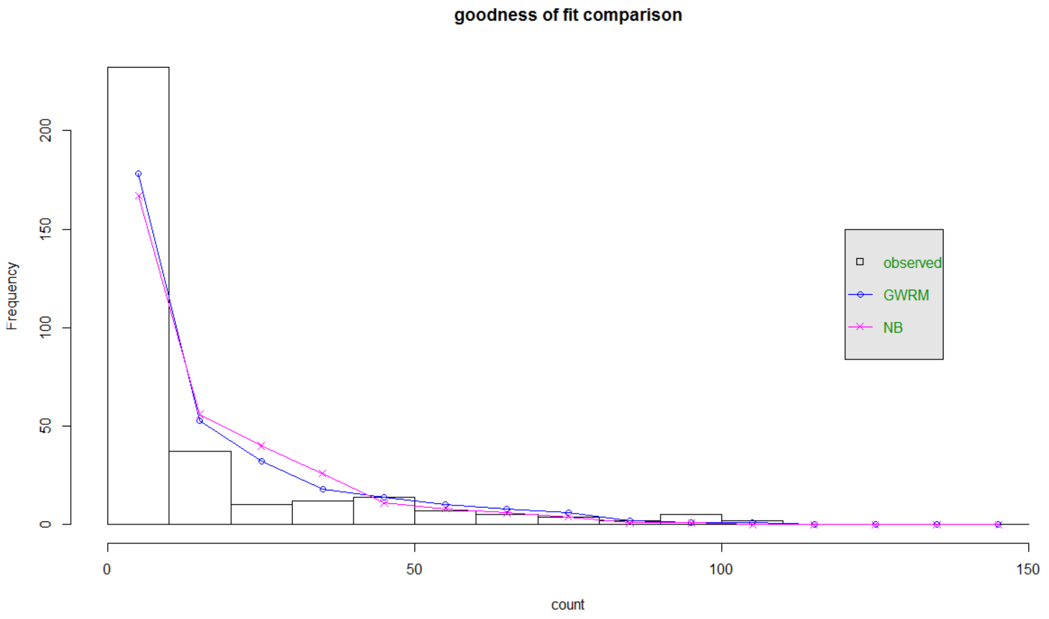

Figure 5.12 Goodness of fit comparison between the NB and GW models for Indiana data

5.2.2.2 Sources of Variance Analysis

Table 5.9 shows the variance of each component for this crash dataset, according to Equation 3.1.

Table 5.9 Variance of each component for the Indiana crash data

Source of variability	**Variance**	**Variance rate**
Randomness	μ_x	$\frac{0.19}{0.55+\mu_x}$
Liability	$1.8235\mu_x$	$\frac{0.355}{0.55+\mu_x}$
Proneness	$5.13\mu_x{}^2$	$\frac{\mu_x}{0.55+\mu_x}$
Total	$(2.8235\mu_x + 5.13\,\mu_x{}^2)$	1

Figure 5.12 shows the fractions of variance attributed to randomness, liability, and proneness for different crash means μ_x which was determined by all of the values of the covariates included in the model. The relationship between the covariate average

daily traffic and the fractions of each component is indicated in Figure 5.13, and the relationship between the covariate friction and fractions of each component when the other variables included in the model are equal to their means (as determined in the data) is illustrated in Figure 5.14.

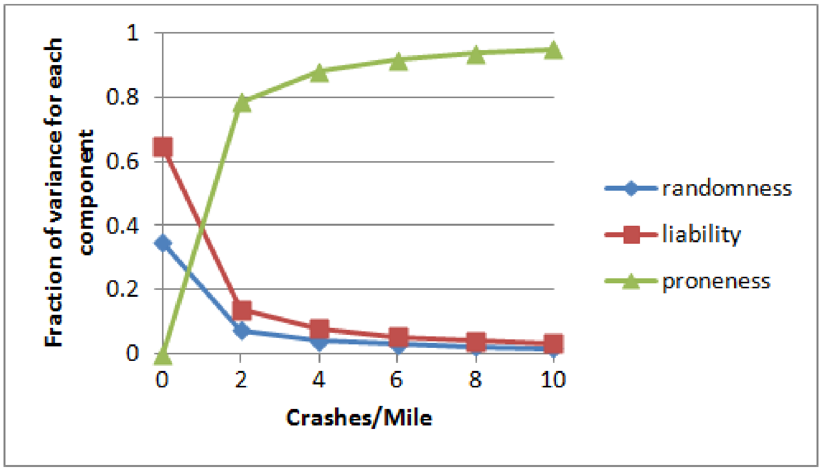

Figure 5.13 Relationship between the fraction of each component and the mean of crashes for Indiana data

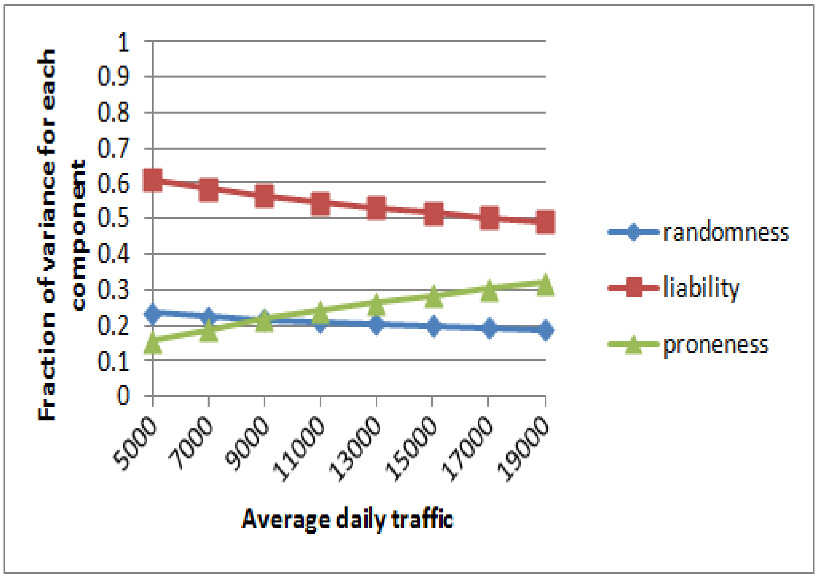

Figure 5.14 Relationship between the fraction of each component and the average daily traffic for Indiana data

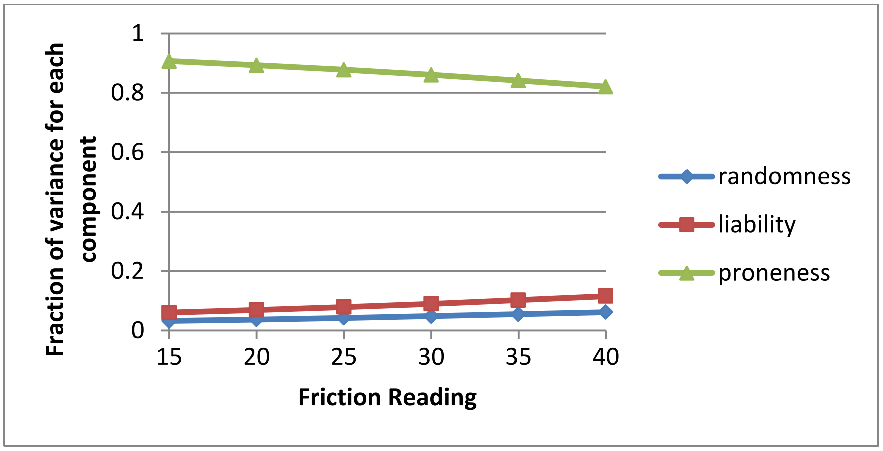

Figure 5.15 Relationship between the fraction of each component and friction reading for Indiana data

It can be seen from Figures 5.12, 5.13, and 5.14 that both randomness and liability decrease from 0.36 to 0.04 and from 0.64 to 0.02, separately, as the mean of crashes on each segment μ_x increases from 0 to 10, which means that both randomness and liability decrease when the ADT and pavement type increases. Both randomness and liability decrease from 0.24 to 0.19 and from 0.61 to 0.48, separately, as the ADT on each segment increases from 5,000 to 19,000. On the other hand, proneness decreases from 0.90 to 0.82 as the friction increases from 15 to 40. Accordingly, both randomness and liability increase from 0.03 to 0.06 and from 0.11, separately, as the friction increases from 15 to 40. The proneness also increases with the presence of interior rumble strips and median barriers. It can be seen that the effect of the friction reading on the fraction of each component is not as significant as is the effect of the ADT.

It should be noted that there are some additional variables included in this model as compared to the model used for the Texas data. Therefore, the total value of the over-dispersion is lower than that of models that include fewer independent geometric or environmental variables. Accordingly, the fraction of variance for randomness is higher when fewer variables are included in the model. The fraction of variance for liability increases and the fraction of variance for proneness decreases when more covariates are included in the model.

5.2.2.3 Covariate Sensitivity Analysis

Figure 5.14 shows the sensitivity of covariates F1 and F2 for both the NB and GW models. It can be seen that these two models have slightly different trends. For both models, a one unit increase in the ADT would result in a significant increase in the logarithm form of the estimated crash when the ADT is below 15,000 veh/day; this increasing rate becomes smaller when the ADT becomes larger. The decrease is more significant for the GW model than for the NB model.

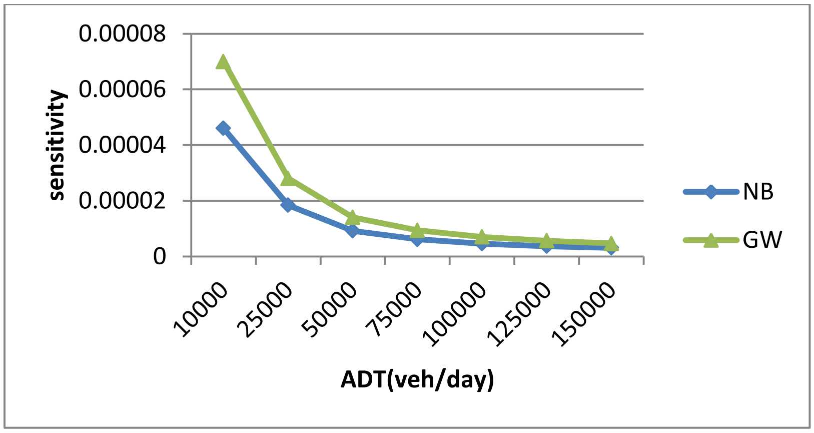

Figure 5.16 Sensitivity analysis of the covariate ADT for the Indiana data

5.2.3 Michigan Crash Data

This dataset is related to single-vehicle crashes that occurred on rural two-lane segments in Michigan State in 2006. This dataset was originally collected for the Federal Highway Administration and has been analyzed by other safety researchers such as Qin et al. (2004) for use in developing crash models. There are many more segments in this dataset, as compared to the above two datasets. Around 70% of all the segments did not include any reported crashes.

Table 5.10 Summary statistics for the Michigan Data

Variable	**Min**	**Max**	**Average(std. dev)**	**Total**
Number of crashes(1 years)	0	61	0.68(1.77)	23168
Annual average daily traffic (AADT)	160	20994	4507.5(3280.6)	-
Segment length(miles)	0.001	54.54	0.18(0.58)	6212
Shoulder width(in feet)	0	24	16.94(5.26)	-
lane width(in feet)	8	15	11.22(0.78)	-
Speed limit (SPEED) (mph)	25	55	52.47(6.39)	-

5.2.3.1 Modeling Estimation

The objective of this example is to compare the performances of the NB-L and GW models on empirical crash data. The parameter-estimating method for this example is the Bayesian method. Within the Bayesian framework of the GW model, as discussed in Chapter 3, the mean response for the number of crashes *y* includes the following formulation:

1.(Y | x) ~ Poisson (λ_x)

2. λ_x ~ Gamma(a_x, v)

3. v ~beta (ρ , k),

It should be noted that one of the most important steps in using Bayesian parameter estimation is to define the priors of all unknown parameters. In this study, normal priors for β, a beta prior for ρ , and a gamma prior for k were used.

On the other hand, within the Bayesian framework of the NB-L regression model, the mean response for the number of crashes y has the following formulation (Lord and Geedipally, 2011):

$P(Y = y, \mu, \phi \mid \varepsilon) = NB(y; \phi, \varepsilon\mu)$

$\varepsilon \sim Gamma(\varepsilon;1+z,\theta)$

$z \sim Bernoulli(z;\frac{1}{1+\theta})$

In this study, normal priors for β , a beta prior for $\frac{1}{1+\theta}$, and a gamma prior for $\frac{1}{\phi}$ were used.

5.2.3.2 Goodness of Fit Analysis

The modeling results of the NB and GW models are presented in this section. Five covariates are considered in this analysis. They include the average daily traffic, lane width, shoulder width, speed, and the length of the segment. The coefficient values of all the covariates in the three models have the same sign and are close to each other. Crashes increase with increases in the AADT, length of segment, and speed in all three models. All three models also show that crashes increase almost linearly with an increase in the length of the segment. It can be seen from the goodness of fit statistics that the GW and NB-L regression models perform much better than the NB model because the values of the DIC, MAD and Pearson χ^2 obtained from the GW and NB-L RM models are smaller than value obtained from the NB model.

Table 5.11 Modeling results for the Michigan data

Variable	NB	NB-L	GW
Intercept	-3.412(0.239)	-3.260(0.193)	-1.233(0.25)
Ln (AADT)	0.426(0.014)	0.424(0.015)	0.496(0.02)
L	0.9571(0.009)	0.961(0.009)	0.9504(0.006)
SW	-0.00009(0.002)	-0.0003(0.002)	-0.019(0.001)
LW	0.0589(0.013)	0.0508(0.011)	0.042(0.002)
SPEED	0.0098(0.002)	0.1024(0.002)	0.0036(0.007)
ρ	-	-	7.53(1.590)
k	-	-	13.35(2.930)
$\alpha = 1/\phi$	0.573	0.102	-
DIC	59354	**56046**	**56497**
MAD	0.651	**0.648**	**0.649**
Pearson χ^2	49911	**44774**	**45826**

5.2.3.3 Sources of Variance Analysis

Table 5.12 shows the variance of each component for this crash dataset according to Equation 3.1. Figure 5.15 shows the fractions of variance attributed to randomness, liability, and proneness for different mean of crashes μ_x .

The relationship between the covariate average daily traffic and the fractions of each component are also indicated in Figure 5.16 when the other variables included in the model are equal to the sample mean values, as observed in the dataset.

Table 5.12 Variance of each component for the Michigan crash data

Source of variability	**Variance**	**Variance rate**
Randomness	μ_x	$\frac{3.71}{13.35+\mu_x}$
Liability	$2.59\mu_x$	$\frac{9.63}{13.35+\mu_x}$
Proneness	$0.27\mu_x^2$	$\frac{\mu_x}{13.35+\mu_x}$
Total	$3.59\mu_x + 0.27\mu_x^2$	1

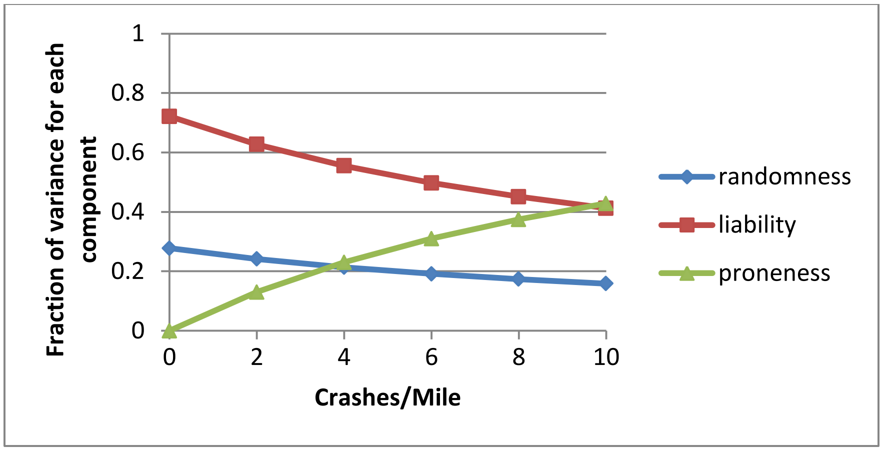

Figure 5. 17 Relationships between the fractions of each component and the mean of crashes for Michigan data

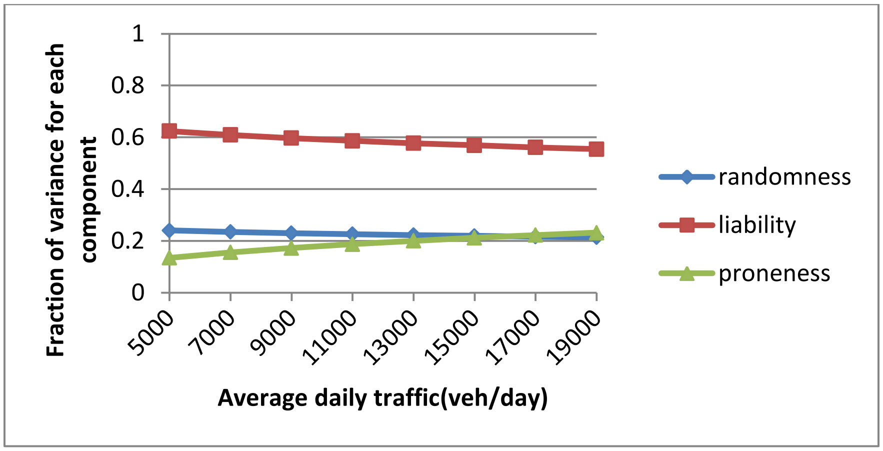

Figure 5. 18 Relationship between the fraction of each component and average daily traffic for Michigan data

It can be seen from Figures 5.15 and 5.16 that randomness and liability decrease from 0.27 to 0.14 and from 0.72 to 0.39, separately, as the mean of crashes on each segment μ_x increases from 0 to 10, which means that both randomness and liability decrease when ADT, length of segment, and speed increases; on the other hand, proneness decreases with an increase in shoulder width.

5.2.3.4 Covariate Sensitivity Analysis

Figure 5.17 shows the sensitivity of covariates F1 and F2 for all of the NB, NB-L, and GW models. It can be seen that these three models exhibit quite similar trends. When the ADT is below 20,000 veh/day, a one unit increase in the ADT would result in a significant increase in the logarithm form of the estimated crash. However, this increasing rate becomes smaller when the ADT becomes larger.

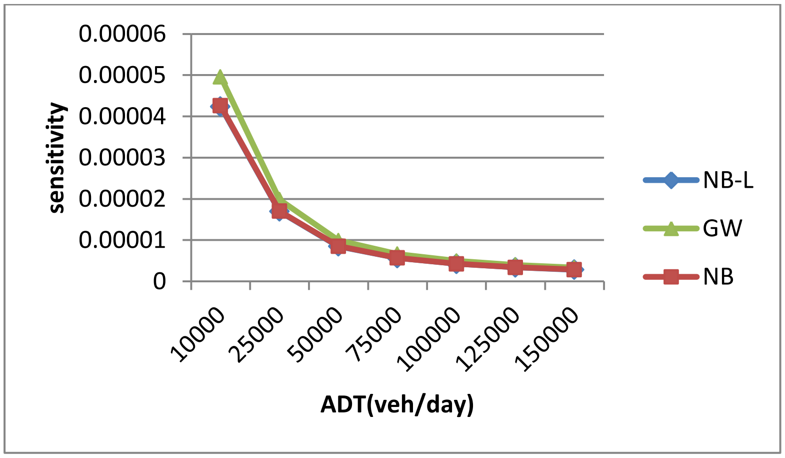

Figure 5. 19 Sensitivity analysis of the covariate ADT for the Michigan data

5.3 Chapter Summary

The objective of this chapter was to apply the generalized regression models to actual vehicle crash data and to demonstrate their ability in discerning the sources of variance within the data. The results of these models were compared with those produced from the standard NB regression model in terms of both goodness of fit and information about sources of variance.

Two kinds of datasets were considered for the application: intersection crash data (see Section 5.1) and segment crash data (see Section 5.2). The applications with these four empirical crash datasets showed that the GW model was a good candidate for characterizing the randomness of crash occurrences and provided useful information,

especially on sources of variations in traffic crashes. Although the GW model did not improve the performance in the goodness of fit aspect for the intersection crash dataset, it provided more detailed information about sources of variance in traffic crashes at intersections by dividing them into three parts; such detailed information could not be detected if we used the NB model. This information is valuable because it will help traffic engineers to better control the variance of traffic crashes by implementing more cost-effective safety countermeasures. The sources of variance are shown to be highly dependent on the traffic flow variable based on the results of all the empirical datasets.
On the other hand, for all three subsequent segment crash datasets, the GW model improved the goodness of fit in addition to providing more valuable information about sources of variance, as compared to the NB model. There are additional variables related to road segments that are included in the Indiana data (as compared to the Texas data) which decreases the difference in results when fitted using both the GW and NB models. It should be noted that the fraction of variance for proneness decreases when more variables related to road segments are included in the model.

Finally, the NB-L model was also applied in the last segment crash dataset within the Bayesian framework in order to provide a high level of completeness of this research. The NB-L model provided a slightly better performance in the goodness of fit aspect in this case, but it is impossible regularly to obtain such a significant amount of detailed information about sources of variance as needed for the GW model. The next chapter presents the application side of the developed GW model in the identification of hotspots.

CHAPTER VI

APPLICATION TO HOTSPOT IDENTIFICATION

The main objective of the previous chapter was to compare the performances of several regression models in empirical crash datasets and accordingly recommend the most appropriate statistical model for a given dataset. The GW regression model was chosen as a more appropriate model than the NB model for highly dispersed crash data. This chapter examines the performance of GW in further by focusing on the application side of the GW regression model in hotspot identification. Although the better performance of the GW model for crash datasets is shown in the previous chapter, it is still necessary to investigate whether this type of model will result in obviously better performance in hotspot identification when compared to the standard NB regression model.

Therefore, the main objective of this chapter is to compare the performances of the GW and NB models in hotspot identification. Section 6.1 introductions the general concept of hotspot identification. Sections 6.2 and 6.3 compare the two models using empirical and simulated data respectively. Section 6.4 summarizes the results of the comparison.

6.1 Hotspot Identification

This section briefly gives an overview of materials related to hotspot identification in traffic safety analysis.

A hotspot also defined as a black spot or a hazardous location can generally be defined as a location such as a roadway segment, intersection, or interchange with a high crash risk (Park, 2010). Various definitions have been used to describe crash risk at a certain location. For example, Hakkert and Mahalel (1978) suggested that a hotspot be defined as a site that has a crash frequency which is significantly higher than expected number given the estimated crash count models. Recently, Elvik (2008) introduced a more reasonable definition of a hotspot that has a higher expected number of crashes compared to other similar locations with the same independent variables.

There are various methods existed have been applied to identify hotspots. The most common and fundamental approach to identifying hotspots is to rank locations based on their actual crash frequencies for a given site. However, this approach can not address the effects of the random characteristic of crash frequency and accordingly cause bias. This kind of bias becomes more significant when the value of the dependent variable is lower, which has been confirmed by some researchers (Miaou and Song, 2005) using simple simulations. Therefore, traffic safety researchers recently prefer to use statistical modeling to addresses the random effects in order to identify the true hot spots more accurately (Miranda-Moreno et al., 2005). They have introduced criteria including “sensitivity” and “specificity” to compare different statistical models in the ability to identify hotspots (Miranda-Moreno, 2006; Elvik, 2008). These criteria can provide information about two types of error including “false positives” which means identifying a non-hotspot site as a hotspot and “false negatives” which may identify a hotspot as a safe site (Park, 2010). These criteria together with some other criteria mentioned in the following section will be used later in this chapter to compare the relative performances of the GW and NB models in identifying hotspots.

Among many ranking functional forms, the following conditional means of crash frequency that are assumed for both the GW and NB models considering the consistence of analysis in Chapter 5 and other researches (Park, 2010):

$$\hat{\mu}_I^{NB} = \exp(X_i \hat{\beta}_{NB}) \quad \text{(NB model)} \tag{6.1}$$

$$\hat{\mu}_I^{GW} = \exp(X_i \hat{\beta}_{GW}) \quad \text{(GW model)} \tag{6.2}$$

6.2 Comparison by Empirical Crash Data

The Texas highway segment data mentioned in the previous chapter were used to compare the differences in ranking orders between the NB and GW models considering the consistence of the modeling results in the previous chapter. The values of $\hat{\mu}_I^{NB}$ and $\hat{\mu}_I^{GW}$ were calculated based on the parameter estimation results in Table 5.5.

Figure 6.1 illustrates the comparison of ranking orders related to the hotspot identification for the two models. For comparison, 300 sites were selected from a list ranked according to values of $\hat{\mu}_I^{GW}$. Smaller values in the ranking order represent higher values in terms of $\hat{\mu}_I^{GW}$, and vice versa. Same ranking rule based on the values $\hat{\mu}_I^{NB}$ was also assigned to these selected sites.

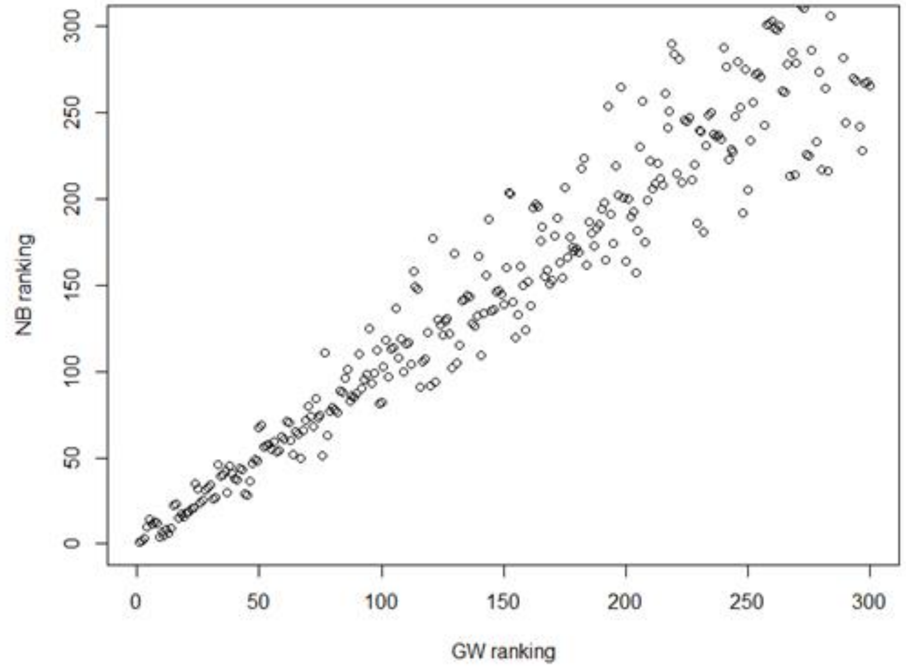

Figure 6.1 Comparison of rankings between the NB and GW models

Figure 6.1 shows that there is a strong positive association between the two rankings. It can also be seen from Figure 6.1 that the difference in rankings becomes larger as the ranking order increases above 100. This research then further compared the ranking orders got from the NB model with those ranked from the GW model in order to figure out the specific differences between the results of the two models (see Figure 6.1).

It is can be seen from the following table 6.1 that fifty sites (m = 50) were selected as hotspots from the GW model; only one site was not included as a hotspot when estimated by the NB model. In addition, five sites for m = 100, 10 sites for m = 200, and 17 sites for m = 300 were not on the hotspot list ranked by the NB model, whereas the list about GW model included them. The percentage deviation used to compare two ranking orders for the number of sites that are different in the two lists of hotspots (Miranda-Moreno et al., 2005) are computed where s is the number of hotspots common in the two lists and m is defined as above. The computation results are shown in Table 6.1 as well. The ranking of the top 50 hotspots are almost the same for both models, but the difference in ranking tends to increase as the number of hotspots selected increases. The ranking results from the GW model may be more reliable than those of the NB model due to a better ability to accommodate over-dispersion.

Table 6.1 Percent deviation of hotspot identification between the NB and GW models

m	50	100	200	300
s	49	95	190	283
% deviation	2%	5%	5%	5.6%

6.3 Comparison using Simulation

It can only be seen from above results that there is some difference in ranking orders between the two models. It is necessary to confirm the conclusion that the ranking results from the GW model may be more reliable than those of the NB model due to a better ability to accommodate over-dispersion. For this reason, this research applies the simulation approach proposed by Miranda-Moreno (2006) to compare the performance of these two models in hotspot identification in further. In simulation, the true hotspot is defined as a site whose expected crash mean is greater than a pre-specified threshold value. Once the true hotspots are identified, it can be considered as the reference to compare the performance of these two models in hotspot identification. The performance evaluation criteria and simulation design are described below (Miranda-Moreno, 2006).

It can be seen in the following table 6.2 that n represents the total number of sites in the set under analysis. The value V and R correspond to the Type I and Type II errors mentioned in section 6.1. It should be noted that the threshold value should be carefully selected to optimize both Type I and Type II errors because these two errors conflict with each other. The lower the Type I error is, the higher the Type II error is. In addition, both errors lead to unnecessary costs. False positives lead to waste in costs related to unnecessary improvements in safety countermeasures. False negatives cause additional costs of traffic crashes in hotspots have not been identified. In this research, another major purpose is accordingly to investigate the effects of threshold values by using two different values.

Table 6.2 Possible outcomes of classification (Miranda-Moreno et al., 2006)

	Number of sites "detected" as non-hotspots	**Number of sites "detected" as hotspots**	
Number of "true" non-hotspots	U	V	N_0
Number of "true" hotspots	R	S	N_1
	n-D	D	n
Where n: the total number of sites in the set under analysis N_0: number of "true" non-hotspots N1: number of "true" hotspots U: number of sites correctly identified as non-hotspots V: number of Type I errors R: number of Type II errors S: number of sites correctly identified as hotspots D: number of sites identified as hotspots			

The following five measures were used as performance criteria to evaluate the relative performances of the two models in detecting the true hotspots (Miranda-Moreno, 2006):

False Discovery Rate (FDR): the ratio of false positives (Type I errors) among all detected hotspots by a model. Smaller values are better.

$$\text{FDR}=\frac{V}{D} \tag{6.3}$$

False Negative Rate (FNR): the ratio of false negatives (Type II errors) among all detected non-hotspots by a model. Smaller values are better.

$$\text{FNR}=\frac{R}{n-D} \tag{6.4}$$

Sensitivity (SENS): the ratio of correctly detected hotspots. Larger values are better.

$$\text{SENS}=\frac{S}{n_1} \tag{6.5}$$

Specificity (SPEC): the ratio of correctly detected non-hotspots. Larger values are better.

$$\text{SPEC}=\frac{U}{n_0} \tag{6.6}$$

Risk (RISK): the ratio of the total number of false positives and false negatives among all the sites under analysis. Smaller values are better.

$$\text{RISK}=\frac{V+R}{n} \tag{6.7}$$

In the simulation design process, this researcher utilized the same covariates and model parameters as the one described in scenario three, discussed in Chapter IV of this research. The crash frequency at each site was assumed to follow the FNP-2 distribution with known parameters in order to examine any potential bias accompanying the use of either the NB or GW model distributions to compare performances. In detail, the datasets were generated by selecting the following values for the parameter: w=0.2 and β_0=2, β_1= 0.5, β_2=-0.5 for the Poisson ($\mu_{i,1}$) distribution and β_0=0, β_1= -0.5, β_2=0.5 for the Poisson (μ_{i2}) distribution.

The simulation was carried out based on the following steps (Park, 2008):

Step 1: The true mean of the crash at site *i* is generated using the following conditional mean functional form:

$$\hat{\mu}_I^{True} = w_1 \exp(X_i \hat{\beta}_1) + w_2 \exp(X_i \beta_2) \tag{6.8}$$

The covariate X_i and the other parameters are defined as in Example 3 in Chapter IV. The data are generated for 1,000 sites.

Step 2: A threshold value *k* was assigned. In this study, two alternative threshold values were analyzed: sample mean and 85th percentile. The following selection rule was applied for each site given the assigned threshold:

1. If $\hat{\mu}_I^{True}$ > k, set H_i=1 and site i ii defined as a “true” hotspot

2. Otherwise, set H_i = 0 and site i is defined as a “non-true” hotspot

Then, summing h_i over n sites results in the total “true” number of hotspots.

Step 3: For each site, simulate the crash frequency based on the method described in Chapter 4.

Step 4: Based on the simulated crash frequency, the model parameters are estimated for the NB and GW models, respectively. The parameter-estimating method follows the one described in Section 4.3.2. This step results in $\hat{\mu}_I^{NB}$ ˆ and $\hat{\mu}_I^{GW}$, as defined in Equations 6.1 and 6.2.

Step 5: Once $\hat{\mu}_I^{NB}$ and $\hat{\mu}_I^{GW}$ are obtained for each site, the following selection rule is applied to identify the "detected" hotspots:

If $\hat{\mu}_I$ > k, set d_i=1 and site i is defined as a "detected" hotspot

Otherwise, set $d_i = 0$ and site i is defined as a "non-detected" hotspot

Summing d_i over n sites results in the total "detected" number of hotspots (D):

$$D=\sum_{i=1}^{n} d_i \tag{6.9}$$

Step 6: At the end of each simulation replication, the five performance criteria (FDR, FNR, SENS, SPEC and RISK) are computed, which are defined in Equations 6.4 to 6.8.

The simulation is replicated 100 times and the average of the 100 replications is used to produce the final results.

6.4 Results

The results of all five performance criteria for the two models are shown in Table 6.3. It should be noted that the results were obtained by using the sample mean value as a threshold value. The average values of five performance criteria for the GW model are all superior to those for the NB model. This simulation confirms the conclusion that the performance of GW model is better than NB model for hotspot identification.

In detail, it can be seen from the results that the false discovery rate (0.38) and false negative rate (0.04) for the NB model are obviously larger than the corresponding values in the GW model, and the sensitivity rate (0.81) and specificity (0.82) for the NB model are smaller than that of the GW model.

Table 6.3 Results of performance criteria measurements when the sample mean is a critical value

Criteria	GW			NB		
	Average	Min	Max	Average	Min	Max
FDR (smaller is better)	0.33	0.12	0.57	0.38	0.21	0.68
FNR (smaller is better)	0.02	0	0.11	0.04	0.01	0.28
SENS (larger is better)	0.97	0.82	1	0.81	0.68	0.98
SPEC (larger is better)	0.84	0.73	1	0.82	0.65	0.91
RISK (smaller is better)	0.12	0.04	0.18	0.16	0.07	0.34

Another simulation was carried out with a different threshold value, the 85th percentile in order to confirm the above conclusion in further. Moreover, this simulation was used to compare the effects of different threshold values on performance. Table 6.4 shows the simulation results from using the 85th percentile threshold values in the sample.

Table 6.4 Results of performance criteria measurements when the 85th percentile is the critical value

Criteria	GW			NB		
	Average	Min	Max	Average	Min	Max
FDR (smaller is better)	0.38	0.12	0.57	0.43	0.22	0.73
FNR (smaller is better)	0.01	0	0.11	0.03	0	0.26
SENS (larger is better)	0.94	0.82	1	0.75	0.63	0.94
SPEC (larger is better)	0.91	0.73	1	0.85	0.67	0.95
RISK (smaller is better)	0.15	0.04	0.18	0.19	0.09	0.38

It is shown in the above table 6.4 that the same conclusion is confirmed and the effects of different threshold values on the performance criteria can be seen from the results. The total error rate decreases when the 85th percentile is used as the criteria compared to the criteria using the sample mean. On the other hand, the FDR and SPEC criteria increase and the FNR and SENS decrease. Thus, using a higher threshold value reduces the number of target hotspots for treatment, which has been pointed out by Park (2010). Therefore, we are more likely to have an increased cost related to the number of non-hotspots misidentified as hotspots if we increase the threshold value, as well as less success in identifying true hotspots. On the meanwhile, we can reduce the costs related to misidentification and increase our ability to detect true non-hotspots. Therefore, a decision on the threshold value should be made by considering the trade-offs between these two costs in order to optimize the limited budget and increase the efficiency for a certain project (Park, 2010).

6.5 Chapter Summary

This chapter first presented a overview of previous research related to hotspot identification in highway safety, and then made a comparison of performance in hotspot identification between the NB and GW models using both empirical and simulated data. The comparison results obtained from the empirical data illustrate a strong positive association between the two rankings for both models and the results from simulated data illustrate the better performance of GW model in hotspot identification compared to NB model. The next chapter focuses on an evaluation of the performance of the GW distribution within a Bayesian framework, in terms of stability and presence of bias for different kinds of datasets especially those characterized by a small sample size and a low sample mean.

CHAPTER VII

SIMULATED ANALYSIS OF BAYESIAN STATISTICS

Small sample size is a common difficulty encountered by traffic safety analysts when applied statistical models because of the limited resources involved in collecting crash data and the variables that influence the number of crashes (Lord and Bonneson, 2005). Many researchers have examined the biasness of the estimators of parameters in the NB model when the data are characterized by a small sample size and low sample mean values. Therefore, the first objective of this chapter is to examine the bias in the parameter estimation of the GW model when the data exhibit a low sample mean and a small sample size. Moreover, it is shown in previous research (Geddipally, 2008) that the prior specifications for the parameters in the statistical model may have a potential influence on the posterior estimation. The other objective of this investigation is to analyze the effects of different prior specifications on the bias. Finally, this chapter will recommend a minimum sample size for applying GW models into crash datasets with different sample means. This recommendation is designed to control unreliable estimations of the posterior mean of the parameters into an acceptable level.

This chapter is divided into three parts. The first part outlines the characteristics of the simulation study. The second part presents the simulation results with different prior specifications and various sample sizes and sample means. The third part provides a summary of the results and brief guidelines for making choices regarding priors and sample sizes.

7.1 Simulation Design

This section briefly describes the simulation study that illustrates the effects of LSM and SSS on the prediction of parameters for GW models. The prior distribution is defined in two different scenarios: a non-informative prior and a weakly-informative prior.

The data were then simulated from the GW model using steps similar to those described in Section 4.2.1.

Step 1: Set sample size, β and the parameters of the GW distribution to the required values.

More specifically, the datasets were generated based on three different mean value scenarios by selecting the values of the parameter shown in Table 7.1.

Table 7.1 True values used for generating the GW variables used for simulation

	High mean	**Moderate mean**	**Low mean**
β_0	2	1	-0.5
β_1	0.5	0.5	0.5
β_2	-0.5	-0.5	-0.5
ρ	3.5	3.5	3.5
k	2.5	2.5	2.5
Sample size	100~1100 (two hundred steps)	100~1100 (two hundred steps)	100~1100 (two hundred steps)

Step 2: Generate two covariates (X_{1i}, X_{2i}) from the standard normal distribution.

Step 3: Calculate the parameter μ_i by using an assumed log-linear function, which is commonly encountered in highway crash analysis $\mu_i = e^{\beta_0 + x^{'}\beta} = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}$.

Step 4: Generate the count variable y_i from the univariate GW distribution UGWD (a_i, k, ρ) where $a_i = \mu_i (\rho - 1)/k$ by using the *rghyper* function in the Suppdist R package.

The MCMC implementation was then used for the model estimation process. Non-informative priors and weakly-informative priors for the two parameters related to variance were utilized for the different scenarios mentioned in the above paragraph. The first was the non-informative prior: $\rho \sim \Gamma(0.01, 0.01), k \sim \Gamma(0.01, 0.01)$.

It has been pointed out by researchers that the above non-informative priors are not good choices in certain cases (Lord and Miranda-Moreno, 2008). The large variance can introduce significant bias, especially for the datasets with low sample mean and small sample size. Therefore, weakly-informative priors have been proposed for use in analyzing vehicle crash data (Washington and Oh, 2006; Miranda-Moreno et al., 2008). Therefore, another commonly used weakly-informative prior was introduced for both parameters of the GW model: $\rho \sim \Gamma(0.5, 0.1), k \sim \Gamma(0.5, 0.1)$. The mean of this prior is equal to 5 and the variance was reduced to 50.

A total of three Markov chains with 25,000 iterations were used initially to check the convergence. A satisfactory convergence was achieved for 25,000 iterations. Then, a single chain with 50,000 iterations and a thinning of 10 were assigned in the Bayesian model estimation process. The first 25,000 iterations were considered to be burn-in samples and only the remaining 25,000 samples were used for estimating the coefficients. The simulation was replicated 100 times for each combination of sample size and parameters. The posterior statistics including posterior mean and standard deviation for each parameter estimation were recorded for each simulation. The bias information and the mean squared error MSE were also calculated to check for the quality of the estimator.

7.2 Simulation Results

This section summarizes the simulation results for both scenarios. The first section summarizes the results of the assumption for non-informative priors for the parameter. The second section gives the results based on the assumption of weakly-informative priors for the dispersion parameters of the GW model.

7.2.1 Non-Informative Priors

The parameter estimation results for the high mean are presented in Tables 7.2 to 7.4 below. These data indicate that as the sample size decreases, the standard deviation increases. For a sample size larger than 300, all of the parameters are accurately estimated and the theoretical value of each parameter is almost equal to its predicted value. For a sample size below 300, there is a little bias to the theoretical value of each parameter and the standard deviation in the prediction becomes larger when compared to a larger sample size.

Table 7. 2 Results of parameters for high mean ($\bar{y}$>5)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
E(β_0)	1.632 (0.543)	1.835 (0.352)	1.936 (0.211)	1.958 (0.069)	2.101 (0.045)	2.085 (0.026)
E(β_1)	0.368 (0.321)	0.467 (0.221)	0.498 (0.185)	0.496 (0.158)	0.504 (0.025)	0.501 (0.015)
E(β_2)	-0.380 (0.421)	-0.452 (0.254)	-0.486 (0.241)	-0.495 (0.196)	-0.497 (0.187)	-0.502 0.126)
E(ρ)	6.885 (3.625)	4.298 (2.123)	3.879 (1.962)	3.968 (0.985)	3.625 (0.756)	3.4 (0.524)
E(k)	3.269 (2.635)	2.236 (1.968)	2.635 (1.658)	2.639 (0.987)	2.469 (0.852)	2.563 (0.425)

Table 7.3 Bias in the estimation of parameters for high mean ($\bar{y}$>5)

Sample size	100	300	500	700	900	1100
β_0	-0.368	-0.165	-0.064	-0.042	0.101	0.085
β_1	0.21	0.1	0.13	0.12	0.04	0.01
β_2	0.18	0.12	-0.08	-0.08	-0.04	-0.03
ρ	3.385	0.798	0.379	0.468	0.125	-0.1
k	0.769	-0.264	0.135	0.139	-0.031	0.063

Table 7.4 MSE in the estimation of parameters for high mean ($\bar{y}$>5)

Sample size	100	300	500	700	900	1100
β_0	0.66	0.39	0.22	0.08	0.11	0.09
β_1	0.38	0.24	0.23	0.20	0.05	0.02
β_2	0.46	0.28	0.25	0.21	0.19	0.13
ρ	4.96	2.27	2.00	1.09	0.77	0.53
k	2.74	1.99	1.66	1.00	0.85	0.43

The simulation results for the medium mean are presented in Tables 7.5 to 7.7. There is a little bias compared to the theoretical values for all the sample sizes, especially when the sample size is below 500. As the sample size decreases, the bias and MSE increase.

Table 7.5 Results of parameters for medium mean (1≤$\bar{y}$≤5)

Sample size	100	300	500	700	900	1100
E(β_0)	0.259 (0.356)	0.356 (0.396)	0.825 (0.265)	1.125 (0.384)	1.116 (0.315)	0.896 (0.125)
E(β_1)	0.215 (0.412)	0.364 (0.365)	0.441 (0.215)	0.469 (0.198)	0.472 (0.058)	0.524 (0.068)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
E(β_2)	-0.187 (0.348)	-0.268 (0.175)	-0.574 (0.169)	-0.468 (0.121)	-0.474 (0.085)	-0.514 (0.067)
E(ρ)	11.854 (5.365)	7.362 (3.625)	5.687 (2.845)	4.984 (1.986)	5.623 (1.587)	4.984 (0.986)
E(k)	1.689 (2.986)	5.359 (1.698)	4.356 (1.657)	2.968 (1.258)	2.869 (1.025)	2.469 (0.365)

Table 7.6 Bias in the estimation of parameters for medium mean ($1\leq\bar{y}\leq5$)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
β_0	-0.741	-0.644	-0.175	0.125	0.116	-0.104
β_1	-0.285	-0.136	-0.059	-0.031	-0.028	0.024
β_2	0.313	0.232	-0.074	0.032	0.026	-0.014
ρ	8.354	3.862	2.187	1.484	2.123	1.484
k	-0.811	2.859	1.856	0.468	0.369	-0.031

Table 7.7 MSE in the estimation of parameters for medium mean ($1\leq\bar{y}\leq5$)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
β_0	0.82	0.76	0.32	0.40	0.34	0.16
β_1	0.50	0.39	0.22	0.20	0.06	0.07
β_2	0.47	0.29	0.18	0.13	0.09	0.07
ρ	9.93	5.30	3.59	2.48	2.65	1.78
k	3.09	3.33	2.49	1.34	1.09	0.37

The simulation results for a low mean are presented in Tables 7.8 to 7.10. These tables exhibit similar characteristics to those shown in the moderate mean scenario. For a sample size below 500, the estimators are highly unreliable. The bias becomes negligible as the sample size increases to be larger than 2,000.

Table 7.8 Results of parameters for low mean ($\bar{y}$<1)

Sample size	**100**	**500**	**1000**	**1500**	**2000**	**2500**	**3000**
E(β_0)	-0.964 (0.698)	-0.756 (0.632)	-0.687 (0.548)	-0.435 (0.421)	-0.487 (0.365)	-0.496 (0.102)	-0.498 (0.062)
E(β_1)	0.324 (0.647)	0.362 (0.425)	0.398 (0.378)	0.412 (0.236)	0.478 (0.178)	0.487 (0.078)	0.503 (0.045)
E(β_2)	-0.962 (0.845)	-0.263 (0.368)	-0.296 (0.174)	-0.396 (0.156)	-0.439 (0.132)	-0.476 (0.078)	-0.511 (0.063)
E(ρ)	23.36 (7.63)	10.269 (5.634)	6.325 (4.213)	4.263 (1.269)	3.258 (0.715)	3.369 (0.636)	3.458 (0.414)
E(k)	6.25 (5.32)	4.269 (2.369)	3.698 (1.745)	3.621 (1.035)	2.368 (0.685)	2.458 (0.541)	2.489 (0.323)

Table 7.9 Bias in the estimation of parameters for low mean ($\bar{y}$<1)

Sample size	**100**	**500**	**1000**	**1500**	**2000**	**2500**	**3000**
β_0	-0.464	-0.256	-0.187	0.065	0.013	0.004	0.002
β_1	-0.176	-0.138	-0.102	-0.088	-0.022	-0.013	0.003
β_2	-0.462	0.237	0.204	0.104	0.061	0.024	-0.011
ρ	19.86	6.769	2.825	0.763	-0.242	-0.131	-0.042
k	3.75	1.769	1.198	1.121	-0.132	-0.042	-0.011

Table 7.10 MSE in the estimation of parameters for low mean ($\bar{y}$<1)

Sample size	100	500	1000	1500	2000	2500	3000
β_0	0.84	0.68	0.58	0.43	0.37	0.10	0.06
β_1	0.67	0.45	0.39	0.25	0.18	0.08	0.05
β_2	0.96	0.44	0.27	0.19	0.15	0.08	0.06
ρ	21.28	8.81	5.07	1.48	0.75	0.65	0.42
k	6.51	2.96	2.12	1.53	0.70	0.54	0.32

7.2.2 Weakly-Informative Priors

This section gives the results for the scenario using the assumption of weakly-informative priors.

The parameter estimation results for the high mean are presented in Tables 7.11 to 7.13. Similar to the results of the non-informative prior scenario, as the sample size decreases the standard deviation increases. All of the parameters are accurately estimated and the theoretical value of each parameter is almost equal to its predicted value for sample sizes greater than 300. It is shown that the posterior statistics including dispersion parameter performs better than under the non-informative prior scenario which is due to the reduced variance in the weakly-informative prior scenario. The difference in the regression parameters is very small for each of the different prior choices.

Table 7.11 Results of parameters for high mean ($\bar{y}$>5)

Sample size	100	300	500	700	900	1100
E(β_0)	1.863 (0.493)	1.837 (0.353)	1.932 (0.213)	1.956 (0.076)	2.103 (0.043)	2.087 (0.023)
E(β_1)	0.376 (0.313)	0.469 (0.223)	0.496 (0.187)	0.494 (0.159)	0.503 (0.024)	0.501 (0.016)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
E(β_2)	-0.375 (0.426)	-0.453 (0.251)	-0.488 (0.224)	-0.497 (0.176)	-0.499 (0.187)	-0.501 (0.131)
E(ρ)	5.671 (1.325)	3.298 (1.131)	3.579 (0.958)	3.568 (0.652)	3.601 (0.326)	3.456 (0.214)
E(k)	3.269 (1.632)	2.636 (1.203)	2.612 (0.958)	2.601 (0.698)	2.569 (0.526)	2.561 (0.312)

Table 7.12 Bias in the estimation of parameters for high mean ($\bar{y}$>5)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
β_0	-0.137	-0.163	-0.068	-0.044	0.103	0.087
β_1	-0.124	-0.031	-0.004	-0.006	0.003	0.001
β_2	0.125	0.047	0.012	0.003	0.001	-0.001
ρ	2.171	-0.202	0.079	0.068	0.101	-0.044
k	0.769	0.136	0.112	0.101	0.069	0.061

Table 7.13 MSE in the estimation of parameters for high mean ($\bar{y}$>5)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
β_0	0.51	0.39	0.22	0.09	0.11	0.09
β_1	0.34	0.23	0.19	0.16	0.02	0.02
β_2	0.44	0.26	0.22	0.18	0.19	0.13
ρ	2.54	1.15	0.96	0.66	0.34	0.22
k	1.80	1.21	0.96	0.71	0.53	0.32

The simulation results for the medium mean are presented in Tables 7.14 to 7.16. The predicted values are slightly misestimated as compared to the theoretical value for the sample sizes below 300; the standard deviation increased as the sample size decreased.

The bias and the value of standard deviation of the posterior means became highly noticeable for the sample size of 100.

Table 7.14 Results of parameters for medium mean (1≤$\bar{y}$≤5)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
E(β_0)	0.254 (0.353)	0.351 (0.392)	0.823 (0.262)	1.121 (0.385)	1.116 (0.313)	0.898 (0.122)
E(β_1)	0.214 (0.415)	0.361 (0.361)	0.443 (0.217)	0.462 (0.191)	0.476 (0.055)	0.521 (0.061)
E(β_2)	-0.183 (0.344)	-0.266 (0.171)	-0.571 (0.174)	-0.464 (0.123)	-0.473 (0.097)	-0.511 (0.072)
E(ρ)	6.343 (1.365)	5.363 (1.325)	5.081 (0.847)	4.321 (0.682)	3.201 (0.321)	3.478 (0.235)
E(k)	4.689 (1.982)	2.301 (1.421)	2.334 (1.012)	2.368 (0.587)	2.444 (0.265)	2.462 (0.187)

Table 7.15 Bias in the estimation of parameters for medium mean (1≤$\bar{y}$≤5)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
β_0	-0.746	-0.649	-0.177	0.121	0.116	-0.102
β_1	-0.286	-0.139	-0.057	-0.038	-0.024	0.021
β_2	0.317	0.234	-0.071	0.036	0.027	-0.011
ρ	2.843	1.863	1.581	0.821	-0.299	-0.022
k	2.189	-0.199	-0.166	-0.132	-0.056	-0.038

Table 7.16 MSE in the estimation of parameters for medium mean (1≤$\bar{y}$≤5)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
β_0	0.83	0.76	0.32	0.40	0.33	0.16
β_1	0.50	0.39	0.22	0.19	0.06	0.06
β_2	0.47	0.29	0.19	0.13	0.10	0.07
ρ	3.15	2.29	1.79	1.07	0.44	0.24
k	2.95	1.43	1.03	0.60	0.27	0.19

The simulation results for the low mean are presented in Tables 7.17 to 7.19. These tables exhibit similar characteristics as those demonstrating the moderate mean scenario. For a sample size below 300, the estimators are unreliable. The bias becomes negligible as the sample size becomes larger than 500.

Table 7. 17 Results of parameters for low mean ($\bar{y}$<1)

Sample size	**100**	**300**	**500**	**700**	**900**	**1100**
E(β_0)	-0.965 (0.701)	-0.758 (0.633)	-0.681 (0.549)	-0.439 (0.422)	-0.491 (0.367)	-0.494 (0.102)
E(β_1)	0.325 (0.648)	0.362 (0.424)	0.398 (0.377)	0.414 (0.236)	0.478 (0.179)	0.489 (0.078)
E(β_2)	-0.964 (0.847)	-0.261 (0.368)	-0.296 (0.171)	-0.397 (0.156)	-0.431 (0.133)	-0.476 (0.078)
E(ρ)	13.02 (5.212)	7.269 (3.214)	3.015 (1.012)	4.063 (0.812)	3.758 (0.412)	3.469 (0.219)
E(k)	4.258 (2.213)	3.269 (2.012)	2.698 (1.254)	2.621 (0.785)	2.367 (0.532)	2.474 (0.213)

Table 7. 18 Bias in the estimation of parameters for low mean ($\bar{y}$<1)

Sample size	100	300	500	700	900	1100
β_0	-0.465	-0.258	-0.181	0.061	0.009	0.006
β_1	-0.175	-0.138	-0.102	-0.086	-0.022	-0.011
β_2	-0.464	0.239	0.204	0.103	0.069	0.024
ρ	9.52	3.769	-0.485	0.563	0.258	-0.031
k	1.758	0.769	0.198	0.121	-0.133	-0.026

Table 7. 19 MSE in the estimation of parameters for low mean ($\bar{y}$<1)

Sample size	100	300	500	700	900	1100
β_0	0.84	0.68	0.58	0.43	0.37	0.10
β_1	0.67	0.45	0.39	0.25	0.18	0.08
β_2	0.97	0.44	0.27	0.19	0.15	0.08
ρ	10.85	4.95	1.12	0.99	0.49	0.22
k	2.83	2.15	1.27	0.79	0.55	0.21

7.3 Results

The results of this study show that similar to the NB model, the GW model is also affected by the low sample mean and small sample size bias. Several conclusions can be made from the results of the simulation study presented above.

First, the parameters of the GW model are estimated more accurately for a larger sample mean ($\bar{y}$>5), as expected. Although the estimated values are close to the theoretical value, the standard deviation of the estimates became large for a sample size of 100. The difference between the minimum and maximum values of the estimates increased as the sample size decreased.

Second, compared to the sample described above, the parameter estimates started to generate obvious bias in the median and low sample means when the sample size was below 300. The estimates are highly unreliable and biased for a low sample mean when the sample size is below 100.

Third, the computational time needed for the MCMC implementation of the GW model was also investigated. Datasets with higher sample means required more computational time for a given number of replications than datasets with a low sample mean. The computational time for the MCMC implementation of the GW model was a little bit larger than NB model but not very significant.

Last but not least, the different definitions of priors for the dispersion parameters of the GW model did not have as much effect on the results of the regression parameter estimation as on the estimation of dispersion parameters. The proper definition of a weakly-informative prior distribution is beneficial to the accuracy of parameter estimation.

Based on the modeling presented above, this research suggests certain guidelines regarding the selection of weakly informative priors and a minimum sample size to use in terms of different sample mean values. The guidelines are shown in Table 7.20. The minimum sample sizes required in each scenario (N =300 for high mean, N =500 for moderate mean, and N =1000 for small mean) were determined according to the bias and MSE associated with each parameter, especially those parameters related to dispersion. The biases related to the regression coefficients were much smaller for different sample sizes and sample means than the biases related to the dispersion parameters.

Table 7.20 Recommended sample size in terms of minimizing bias (weakly informative priors)

Sample mean	Recommended Minimum Sample Size
High($\bar{y}$>5)	300
Moderate(1≤$\bar{y}$≤5)	500
Low($\bar{y}$<1)	1000

7.4 Chapter Summary

This chapter evaluated the performance of GW distribution in terms of stability and presence of bias for data with different sample sizes and sample means. The effects of prior distributions were also investigated. It is shown that the GW model was also affected by low sample mean and small sample size bias. The estimates were highly unreliable and biased for the low sample mean when the sample size was below 100.The proper definition of weakly-informative prior distribution was necessary to improve the accuracy of parameter estimation. Finally, the selection of priors and the minimum sample size to use given different sample mean values was also presented.

CHAPTER VIII

CONCLUSIONS AND FUTURE RESEARCH

As one of the major analysis methods, statistical models play an important role in traffic safety analysis. There has been considerable work in the traffic safety literature related to the development of statistical models for analyzing motor vehicle crashes. A common difficulty associated with the modeling of crash data is known as the overdispersion. It is a serious problem and has been addressed in a variety ways. One of the most commonly used methods to address this is with the NB model. However, factors that affect the extra variation are often unknown to researchers, and traditional models, such as the NB model, cannot adequately capture the nature of the dispersion found in crash data.

Given the limitations of the NB regression model for addressing the source of overdispersion of crash data, this research examined an alternative model formulation that could be used for capturing the source of extra variability through the use of the GW regression model. To evaluate its performance, GW regression models were estimated using both simulated and empirical crash datasets, and the results were compared to the NB regression model as well as to the recently introduced NB-L model. Their relative performances were also examined in terms of hotspot identification. Finally, bias properties of the choice of prior distributions for parameters in GW regression model were characterized, and guidelines on the choice of priors and the summary statistics to use were presented for different datasets.

This chapter is divided into two parts. The summary and conclusions of the research are presented in Section 8.1 and future research is discussed in Section 8.2.

8.1 Summary of Work

This section briefly summarizes the major contributions of this research.

In Chapter II, various crash count models addressing both overdispersion and underdispersion crash data have been described and a brief discussion about strengths and weaknesses of existing models was presented. The chapter also described the two estimation methodologies that can be used for estimating the coefficients of regression models.

In Chapter III, the methodology of the univariate GW distribution and the corresponding GW model based on the distribution were presented. In addition, the basic parameter estimation methods applied to a GW model within both the maximum likelihood and the Bayesian framework were developed. Finally, the relationship between the GW distribution and NB distribution was described. It can be inferred that the NB models are nested in the GW.

In Chapter IV, several conclusions were presented using simulated datasets. At first, simulated datasets were used to illustrate how GW regression models can provide good numerical approximations and the poor capability of the traditional NB model when the underlying distribution comes from GW distribution. The effects of sample mean and sample size on the goodness of fit were also examined.

Next, simulated datasets were used to illustrate the appropriateness of the GW regression model specifications when the data were actually generated from a NB distribution. The effects of sample size and degree of dispersion on the GOF were also examined. Moreover, simulated datasets were generated by using a two- component finite mixture of Poisson distribution and a two-component finite mixture of NB distribution in order to examine potential bias with either NB distribution or GW distribution to compare performance. It was shown that GW fitted better than the NB model for highly dispersed datasets generated from two heterogeneous populations. In sum, the GW model was a good candidate model in all four scenarios.

In chapter V, the applications with four empirical crash datasets showed that the GW was a good candidate model to characterize the randomness of crash occurrence. It provided useful information, especially on source of variation of traffic crashes. Although the GW model did not improve the performance in goodness-of-fit aspect for the intersection crash dataset, it provided more detail information about source of variance of traffic crashes at intersections by dividing into three parts, which could not be detected using the NB model. This information is valuable because it will help traffic engineers to better control the variance of traffic crashes by implementing more cost-effective safety countermeasures.

On the other hand, for all three subsequent segment crash datasets, the GW model improved the goodness-of-fit, in addition to providing more valuable information about source of variance compared to the NB model. It is worth noting that the NB-L model was also applied to the last segment crash dataset within a Bayesian framework for the better completeness of this research. It is noted that the NB-L model provided a little better performance in the goodness-of-fit aspect in this case, but it is not possible to provide as much detailed information about source of variance compared to the GW model.

In the next chapter VI, the GW model was applied to compare performance in hotspot identification with the NB model. Five different performance criteria were used in the study. It showed that the average values of five performance criteria for the GW model are all superior to those for the NB model because of the better model specifications to account for variance.

The effects of different threshold values on the performance criteria were also investigated in this research. Compared to the criteria using sample mean, the total error rate indicated by RISK decreased when the 85th percentile was used as the criteria. However, the FDR and SPEC criteria increase and the FNR and SENS decreases\. It is shown that using a higher threshold value reduces the number of target hotspots for treatment. Accordingly it is more likely to have an increased number of false positives and less success in detecting true hotspots as the threshold value increases. At the same time, one can reduce the possibility of misidentifying hotspots as non-hotspots and increase the ability to detect true non-hotspots.

Finally, in chapter VII, this research investigated the effects of different sample sizes and sample means especially under low sample mean and small sample size scenario. A simulation study was used to indicate the effects on bias properties of the GW model at this scenario. The effect of the different prior distributions for the parameters of GW model on the posterior estimation and biases of the parameters was also investigated. Two different priors, beta prior and gamma prior were examined. Based on the simulation results, general guidelines were also provided about the choice of priors and the summary statistics to use for different sample sizes and sample mean values. The minimum recommended sample sizes for each sample-mean category were N =300 for high mean, N =500 for moderate mean, and N =1500 for small mean. It was found that there was no major difference in the prediction of the estimates and the standard deviations with the change in prior distribution.

In conclusion, this research developed GW model applied in traffic safety with both MLE and Bayesian parameter estimation methods and examined the superior performance of this kind of model in three major aspects: goodness of fit, source of variance of traffic crashes and the ability to investigate hotspots. This kind of model will be promising, especially for researchers to investigate the source of variance of traffic crashes in the future.

8.2 Recommendations and Future Research

Highway safety researchers should consider using the alternate GW model if the crash datasets are overdispersed and the data are not well suited to a NB regression model by checking the sample mean and variance. In these cases, the GW model specification could be a good candidate model for this dataset.

Even if the goodness-of-fit performance of the NB model is satisfactory, it could still be meaningful to apply the GW model for analyzing the crash dataset. The model is able to provide at least the same performance on goodness of fit for overdispersed data but more valuable information about the source of variance.

The following are some directions for future study:

The EB method is now commonly used in highway safety analysis for hotspot identification. This research can be extended by developing an EB modeling framework for the GW model.

There are several other models such as random parameter models also have been used to describe sources of dispersion. The difference between these models with GW is meaningful to be investigated and compared in order to understand the source of variance of traffic vehicle crashes better.

In Chapter VII, only two kinds of prior information were discussed and compared. It would probably be useful to investigate some better definitions of prior information for the parameters of the GW model, in order to improve the accuracy of parameter estimation.

There have been some discussions about the assumption related to the covariate-dependent dispersion for NB models. Similar to the NB model with a varying dispersion parameter, further research should be conducted to examine the effect of a covariate-dependent parameter with the GW model.

REFERENCES

America Automobile Association (AAA) 2011. Crash vs Congestion-what is the cost to society. Cambridge Systematics http://newsroom.aaa.com/wp-content/uploads/2011/11/2011_AAA_CrashvCongUpd.pdf

Anastasopoulos, P., Mannering, F., 2009. A note on modeling vehicle-accident frequencies with random-parameters count models. Accident Analysis and Prevention 41(1), 153-159

Cameron, A., Trivedi, P.K., 1998. Regression Analysis of Count Data. In: Econometric Society Monographs, vol. 30. Cambridge University Press, Cambridge.

Cheng, L., S.R. Geedipally, and D. Lord 2012 Examining the Poisson-Weibull Generalized Linear Model for Analyzing Crash Data. Safety Science, Vol. 54, pp. 38-42.

Clark, S.J., Perry, J.N., 1989. Estimation of the negative binomial parameter by maximum quasi-likelihood. Biometrics 45, 309-316.

Elvik, R., 2000. How much do road accidents cost the national economy. Accident Analysis and Prevention 32, 849-851.

Fruwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer Series in Statistics, Springer, New York.

Geedipally, S.R., 2008. Examining the Application of Conway-Maxwell Poisson Model for Analyzing Traffic Crash Data. Ph.D. Dissertation, Department of Civil Engineering, Texas A&M University, College Station, Texas.

Geedipally, S.R., Lord, D. 2008 Effects of the varying dispersion parameter of Poisson-gamma models on the estimation of confidence intervals of crash prediction models. Transportation Research Record 2061, 46-54

Geedipally, S.R., S. Patil, and D. Lord 2010 Examining Methods for Estimating Crash Counts According to their Collision Type. Transportation Research Record 2165, pp. 12-20.

Geedipally, S.R., and D. Lord 2011 Examining the Crash Variances Estimated by the Poisson-Gamma and Conway-Maxwell-Poisson Models. Transportation Research Record 2241, pp. 59-67.

Geedipally, S.R., D. Lord, S.S. Dhavala 2012 The Negative Binomial-Lindley Generalized Linear Model: characteristics and Application using Crash Data. Accident Analysis & Prevention, Vol. 45, No. 2, pp. 258-265.

Gelfand, A. E., and Smith, A. F. M. 1990, Sampling-Based Approaches to Calculating Marginal Densities, Journal of the American Statistical Association, 85, 398409.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis (Second Edition). Chapman & Hall, London.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (6), 721-741.

Ghitany, M. E., Atieh, B., Nadarajah, S. 2008. Lindley distribution and its application. Mathematics and Computers in Simulation, 78(4), 493506.

Gilks, W.R., Rivhstfdon, S. and Spiegelhalter, D.J 1996 Markov Chain Monte Carlo in Practice. London:Chapman and Hall.

Greenwood, M. and Yule, G. U. 1920. An enquiry into the nature of frequency distributions representative of multiple happenings, with special reference to multiple attacks of disease or repeated accidents. J.R. Statist. Soc., 83, 255-279.

Guo, J.Q., Trivedi, P.K., 2002. Flexible parametric models for long-tailed patent count distributions. Oxford Bulletin of Economics and Statistics 64 (1), 63-82.

Hakkert, A.S, Mahalel, D., 1978. Estimating the number of accidents at intersections from knowledge of the traffic flow on the approaches. Accident Analysis and Prevention 10 (1), 69-79.

Hauer, E., 1997. Observation Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Elsevier Science Ltd.,Oxford.

Hauer, E., 2001. Overdispersion in modeling accidents on road sections and in empirical Bayes estimation. Accident Analysis and Prevention 33 (6), 799-808.

Hilbe, J.M., 2007. Negative Binomial Regression. Cambridge University Press, Cambridge, UK.

Irwin, J.O., 1968. The GW distribution applied to accident theory. Journal of the Royal Statistical Society. Series A 131 (2), 205_225.

Jae Myung., 2002. Tutorial on maximum likelihood estimation Journal of Mathematical Psychology 47 (2003) 90–100

John H. Mathews and Kurtis K. Fink 2004 Numerical Methods Using Matlab, 4th Edition

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

Land K.C., McCall, P.L., Nagi, D.S., 1996. A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models. Sociological Methods and Research 24 (4), 387-442.

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

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 and Prevention 38 (4), 751-766.

Lord, D., Bonneson, J.A., 2007. Development of accident modification factors for rural frontage road segments in Texas. Transportation Research Record 2023, 20-27.188

Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., van Schalkwyk, I.,Ivan, J.N., Lyon, C., Jonsson, T., 2009. Methodology for Estimating the Safety Performance of Multilane Rural Highways. NCHRP Web-Only Document 126,

Lord, D., Park, P.Y-J., 2008. Investigating the effects of the fixed and varying dispersion parameters of PG models on empirical Bayes estimates. Accident Analysis and Prevention 40 (4), 1441-1457.

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

Lord, D., Washington, S.P., Ivan J.N., 2007. Further notes on the application of zero Accident Analysis & Prevention, Vol. 39, No. 1, pp. 53-57.

Lord, D., and F. Mannering 2010 The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. Transportation Research - Part A, Vol. 44, No. 5, pp. 291-305.

Malyshkina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching Negative Binomial models: an application to vehicle accident frequencies. Accident Analysis and Prevention 41 (2), 217-226. 189

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21 (6), 1087-1092.

Miaou, S-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transportation Research Record 1840, 31-40.

Miranda-Moreno, L.F., Fu, L., Saccomanno, F.F., Labbe, A., 2005. Alternative risk models for ranking locations for safety improvement. Transportation Research Record 1908, 1-8.

Miranda-Moreno, L.F., 2006. Statistical models and methods for the identification of hazardous locations for safety improvements. Ph.D. Dissertation, Department of Civil Engineering, University of Waterloo, Canada.

Miranda-Moreno, L.F., Lord, D., Fu, L., 2008. Bayesian road safety analysis: incorporation of past experiences and effect of hyper-prior choice. In: Proceedings of TRB 87th Annual Meeting Compendium of Papers DVD, Transportation Research Board, Washington, D.C.

Mitra, S., Washington S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39 (3), 459-468.

Newbold, E. M. 1925. A contribution to the study of the human factor in the causation of accidents. Industrial Health Research Board Report, No. 3.

Newbold, E. M 1927. Applications of the statistics of repeated events, particularly to industrial accidents. J.R. Statist. Soc., 90, 487-547

NHTSA, 2010. National Highway Traffic Safety Administration, National Center for Statistics and Analysis, U.S. Department of Transportation, Washington, D.C. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook

Oh, J., Lyon, C., Washington, S.P., Persaud, B.N., Bared, J., 2003. Validation of the FHWA crash models for rural intersections: lessons learned. Transportation Research Record 1840, 41–49.

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

Park, B.-J., Lord, D., 2008. Adjustment for the maximum likelihood estimate of the Negative Binomial dispersion parameter. Transportation Research Record 2061, 9-19.

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

Park, B.-J., D. Lord, and J. Hart (2010) Bias Properties of Bayesian Statistics in Finite Mixture of Negative Regression Models for Crash Data Analysis. Accident Analysis & Prevention, Vol. 42, No. 2, pp. 741-749.

Park, B.-J., 2010. Application of the Finite Mixture Model for Vehicle Crash Data Analysis. Ph.D. Dissertation, Department of Civil Engineering, Texas A&M University, College Station, Texas.

Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1-6.

Persaud, B.N., Lord, D., Palminaso, J., 2002. Issues of calibration and transferability in developing accident prediction models for urban intersections. Transportation Research Record 1784, 57-64.

Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis & Prevention 36 (2), 183–191.

Ramaswamy, V., Anderson, E.W., DeSarbo, W.S., 1994. A disaggregate negative binomial regression procedure for count data analysis. Management Science 40 (3), 405-417.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A.J., Olmo-Jiménez, M.J., 2008. A new generalization of the waring distribution. Computational Statistics and Data Analysis 51 (12), 6138_6150.

Rodríguez-Avi, J., A. Conde Sánchez, A.J. Sáez Castillo, M.J. Olmo Jiménez, A. M. Martínez Rodríguez (2009). A GW model for count data. Computational Statistics & Data Analysis 53 3717-3725

Saha, K., Paul, S., 2005. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. Biometrics 61 (3), 179–185.

SAS Institute Inc., 2002. Version 9 of the SAS System for Windows. Cary, NC. Scaccia, L., Green, P.J., 2003. Bayesian growth curves using normal mixtures with nonparametric weights. Journal of Computational and Graphical Statistics 12 (2), 308-331.

Shankar, V.N., Albin, R.B., Milton, J.C, Mannering, F.L., 1998. Evaluating median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effects Negative Binomial model. Transportation Research Record 1635, 44-48.

Shankar, V.N., Milton, J., Mannering, F., 1997. Modeling accident frequency as zeroaltered probability process: an empirical inquiry. Accident Analysis and Prevention 29 (6), 829-837.

Shankar, V.N., Ulfasson, G.F., Pendyala, R.M., Nebergall, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Science 41(7), 627-640.

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

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

Spiegelhalter, D.J., Thomas, A., Best, N.G., Lun, D., 2003. WinBUGS Version 1.4.1 User Manual. MRC Biostatistics Unit, Cambridge.

Stein, G., Zucchini, W., and Juritz J., 1987. Parameter Estimation for the Sichel Distribution and Its Multivariate Extension. Journal of the American Statistical Association 82 (399), 938–944.

Tanner, T., Wong, W., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82 (398), 528-549.

Tong, J., Lord, D., 2007. Investigating the application of beta-binomial models in highway safety. Canadian Multidisciplinary Road Safety Conference XVII, June 3-8, 2007, Montreal.

Wang, P., Cockburn, I.M., Puterman, M.L., 1998. Analysis of patent data: a mixed Poisson regression model approach. Journal of Business and Economic Statistics 16 (1), 27-41.

Wang, W., Famoye, F., 1997. Modeling household fertility decisions with generalized Poisson regression. Journal of Population Economics 10 (3), 273-283.

Warton, D.I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16 (3), 275-289.

Washington, S., Oh, J., 2006. Bayesian methodology incorporating expert judgment for ranking countermeasure effectiveness under uncertainty: Example applied to at grade railroad crossings in Korea. Accident Analysis and Prevention 38 (2), 234-247.

Washington, Simon, Cole, Robert J., & Herbel, Susan B. (2011) European advanced driver training programs : reasons for optimism. *IATSS Research*, *34*(2), pp. 72-79.

Wedagama, D.M.P, 2006. The Relationship between Urban Land Use and Non Motorised Transport Accidents and Casualties, PhD Thesis, Newcastle University

World Health Orgnization (WHO), 2009. Global Status Report On Road Safety http://www.who.int/violence_injury_prevention/road_safety_status/2009/en/

World Health Orgnization (WHO), 2012. World Health Statistics Report http://www.who.int/gho/publications/world_health_statistics/2012/en/

Xekalaki, E., 1983. The univariate GW distribution in relation to accident theory: Proneness, spells or contagion. Biometrics 39 (4), 887_895.

Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural networks: an empirical analysis. Accident Analysis and Prevention 39 (5), 922-933.

Zamani, H. and N. Ismail, 2010. Negative Binomiallindley distribution and its application. Math. Statis 6: 4-9. ISSN: 1549-3644. DOI: 10.3844/jmssp.2010.4.9

Zhang, Y., Ye, Z., Lord, D., 2007. Estimating the dispersion parameter of the Negative Binomial distribution for analyzing crash data using a bootstrapped maximum likelihood method. Transportation Research Record 2019, 15-21.

APPENDIX A

R and WINBUGS CODES FOR GW MODEL

This appendix provides codes for generating the GW, NB and FMP-2 random variables and simulated datasets. The implementation of the MLE and Bayesian method for GW model is also provided.

1. Example of generating GW variates and MLE for GW and NB model

```
Library ("GWRM")
Rho<-3.5
k<-2.5
N=100
b=matrix(c(-0.5,0.5,-0.5), nrow=3, ncol=1)
X=cbind(rep(1,N),rnorm(N,mean=0,sd=1),rnorm(N,mean=0,sd=1))
offset=c(rep(0,N))
xbeta=X%*%b
m=exp(xbeta)
a<-m*(rho-1)/k
y<-rghyper(N,-k,-a,rho-1)
mydata=data.frame(y,X)
fit<-GWRM.fit(y~X[ ,2]+X[ ,3],data=mydata)
GWRM.display(fit)
m1<-glm.nb(y~X[ ,2]+X[ ,3],data=mydata)
```

2. Example of generating NB variates and MLE for GW and NB model

```
N=100
b=matrix(c(1,0.5,-0.5), nrow=3, ncol=1)
X=cbind(rep(1,N),rnorm(N,mean=0,sd=1),rnorm(N,mean=0,sd=1))
offset=c(rep(0,N))
```

```
xbeta=X%*%b
m=exp(xbeta)
y<-rnbinom(N,mu=m,size=2)
mydata=data.frame(y,X)
m1<-glm.nb(y~X[ ,2]+X[ ,3],data=mydata)
fit<-GWRM.fit(y~X[ ,2]+X[ ,3],data=mydata)
GWRM.display(fit)
```

3. Example of generating FMP-2 variates and MLE for GW,NB and FMP-2 model

```
library("flexmix")
N=100
b=matrix(c(2,-0.5,0.5,0,-0.5,0.5), nrow=3, ncol=2)
X=cbind(rep(1,N),rnorm(N,mean=0,sd=1),rnorm(N,mean=0,sd=1))
offset=c(rep(0,N))
xbeta=X%*%b
m=exp(xbeta)
w=rbinom(N, 1, .2)
y=w*rpois(N,m[,1])+(1-w)*rpois(N,m[,2])
mydata=data.frame(y,X)
m1<-glm.nb(y~X[ ,2]+X[ ,3],data=mydata)
fit<-GWRM.fit(y~X[ ,2]+X[ ,3],data=mydata)
GWRM.display(fit)
M2 <- flexmix(y~X[ ,2]+X[ ,3],data=mydata, k = 2, model = FLXMRglm(family = "poisson"))
```

4. Example of generating FMNB-2 variates and MLE for GW,NB model

```
library("flexmix")
N=1000
b=matrix(c(2,-0.5,0.5,0,-0.5,0.5), nrow=3, ncol=2)
```

```
X=cbind(rep(1,N),rnorm(N,mean=0,sd=1),rnorm(N,mean=0,sd=1))
offset=c(rep(0,N))
xbeta=X%*%b
m=exp(xbeta)
w=rbinom(N, 1, .2)
y=w*rnbinom(N,mu=m[,1],size=5)+(1-w)*rnbinom(N,mu=m[,2],size=5)
mydata=data.frame(y,X)
m1<-glm.nb(y~X[ ,2]+X[ ,3],data=mydata)
fit<-GWRM.fit(y~X[ ,2]+X[ ,3],data=mydata)
GWRM.display(fit)
Summary (m1)
```

5. Example of generating GW variates and Bayesian estimation for GW model

```
Library (R2WinBUGS)
Rho<-3.5
k<-2.5
N=1000
b=matrix( c(1.5,0.5,-0.5), nrow=3, ncol=1)
X=cbind(rep(1,N),rnorm(N,mean=0,sd=1),rnorm(N,mean=0,sd=1))
offset=c(rep(0,N))
xbeta=X%*%b
m=exp(xbeta)
a<-m*(rho-1)/k
y<-rghyper(N,-k,-a,rho-1)
mydata=data.frame(y,X)
J<-nrow(mydata)
y<-mydata$y
x2<-mydata$X2
x3<-mydata$X3
```

```
crash.data<-list("J","y","x2","x3")
crash.inits<-function()
list(b=c(1.5,0.5,-0.5),rho=3.5,k=2.5)
crash.parameters<-c("b","rho","k")
crash.sim<-bugs(crash.data,crash.inits,crash.parameters,model.file="C:/Desktop/winbug
sim/simGW.bug",n.chains=3,n.iter=50000,n.burnin=25000,bugs.directory="C:
/Desktop/studyME/WINBUGS/WinBUGS14",debug=TRUE)
print(crash.sim,digits.summary=3)
model
{for (i in 1:1000)
   {error[i] ~ dgamma(phi1[i],phi1[i])
      phi1[i]~dbeta(rho,k)
      mu[i]<- exp(X%*%b)*error[i]
      y[i]~dpois(mu[i])
     }
   for (i in 1:3){
b[i]~dnorm(0,1.0E-6)
}
rho~dgamma(0.01,0.01)
k~dgamma(0.01,0.01)
}
```

APPENDIX B

FIGURES ABOUT RELATIONSHIP BETWEEN FRACTION OF EACH COMPONENT AND EACH INDEPENDENT VARIABLE IN GW MODELS

Texas crash data

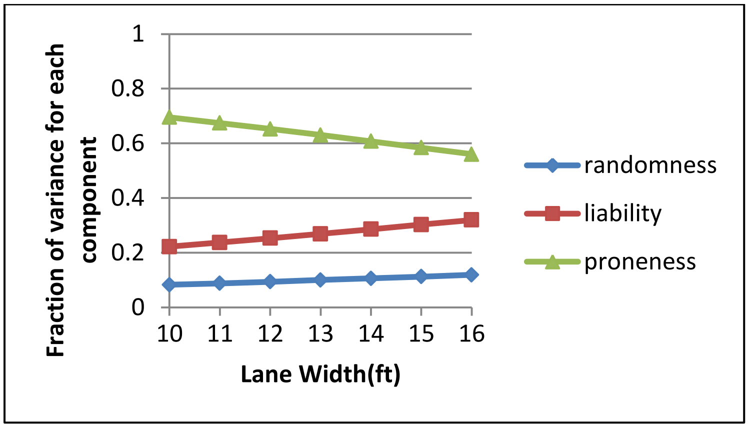

Figure B.1 Relationship between fraction of each component and lane width

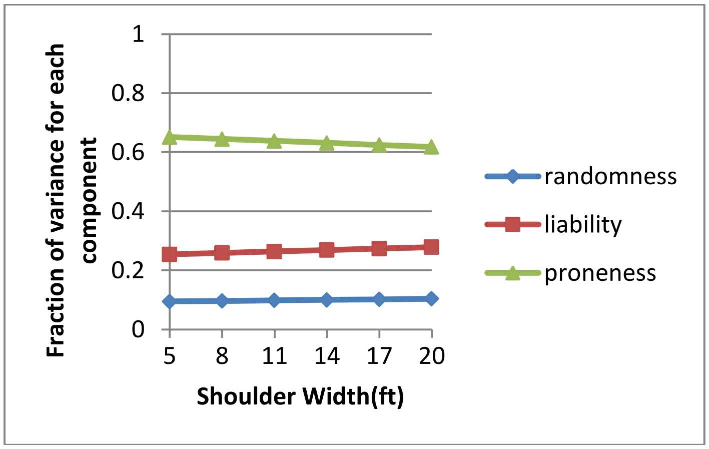

Figure B.2 Relationship between fraction of each component and shoulder width

Indiana crash data

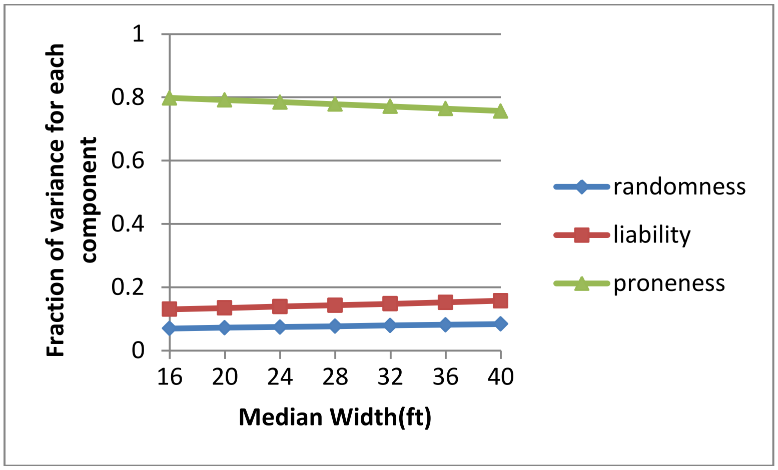

Figure B.3 Relationship between fraction of each component and median width

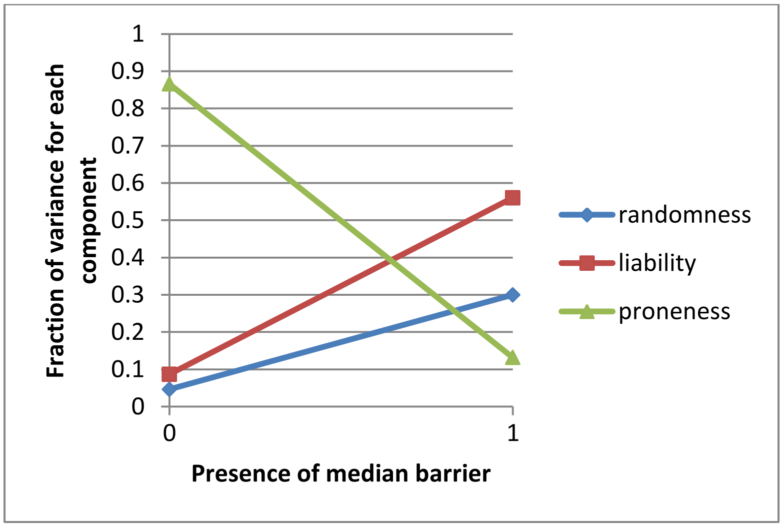

Figure B.4 Relationship between fraction of each component and Presence of median barrier

Michigan crash data

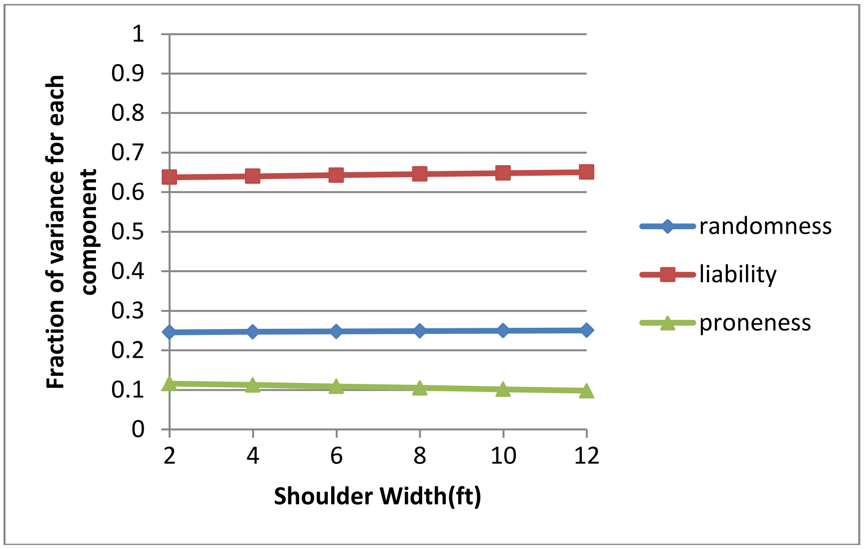

Figure B.5 Relationship between fraction of each component and shoulder width

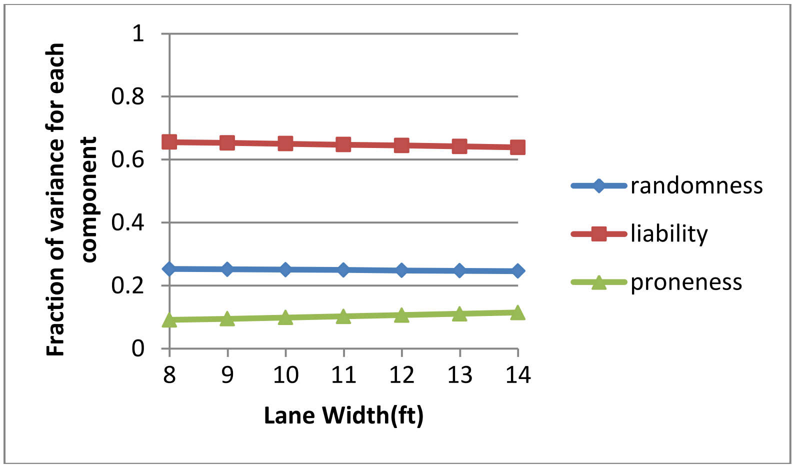

Figure B.6Relationship between fraction of each component and lane width

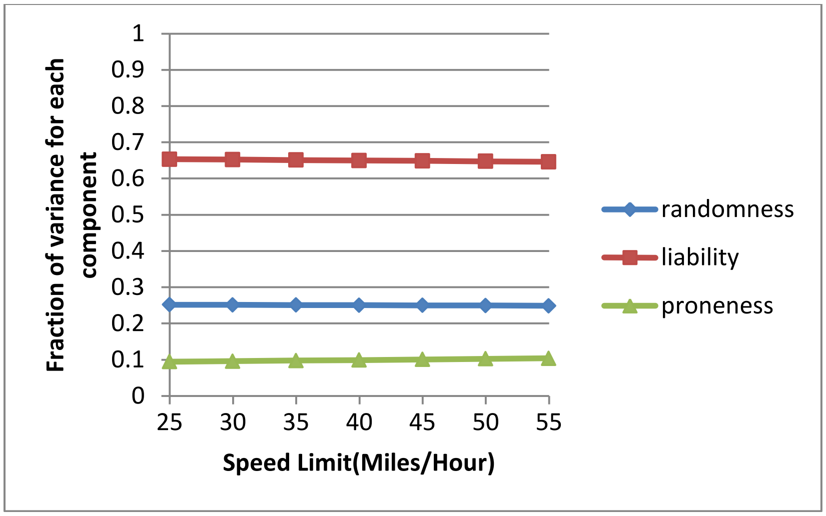

Figure B.7 Relationship between fraction of each component and speed limit