EXAMINING THE APPLICATION OF CONWAY-MAXWELL-POISSON MODELS FOR ANALYZING TRAFFIC CRASH DATA

A Dissertation

by

SRINIVAS REDDY GEEDIPALLY

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

December 2008

Major Subject: Civil Engineering

EXAMINING THE APPLICATION OF CONWAY-MAXWELL-POISSON MODELS FOR ANALYZING TRAFFIC CRASH DATA

A Dissertation

by

SRINIVAS REDDY GEEDIPALLY

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,	Seth Guikema
	Jeffrey Hart
	Samiran Sinha
Head of Department,	David Rosowsky

December 2008

Major Subject: Civil Engineering

ABSTRACT

Examining the Application of Conway-Maxwell-Poisson Models for Analyzing Traffic Crash Data. (December 2008)
Srinivas Reddy Geedipally, B.E., Osmania University;
M.Sc., Linköpings University
Chair of Advisory Committee: Dr. Dominique Lord

Statistical models have been very popular for estimating the performance of highway safety improvement programs which are intended to reduce motor vehicle crashes. The traditional Poisson and Poisson-gamma (negative binomial) models are the most popular probabilistic models used by transportation safety analysts for analyzing traffic crash data. The Poisson-gamma model is usually preferred over traditional Poisson model since crash data usually exhibit over-dispersion. Although the Poisson-gamma model is popular in traffic safety analysis, this model has limitations particularly when crash data are characterized by small sample size and low sample mean values. Also, researchers have found that the Poisson-gamma model has difficulties in handling under-dispersed crash data. The primary objective of this research is to evaluate the performance of the Conway-Maxwell-Poisson (COM-Poisson) model for various situations and to examine its application for analyzing traffic crash datasets exhibiting over- and under-dispersion. This study makes use of various simulated and observed crash datasets for accomplishing the objectives of this research.

Using a simulation study, it was found that the COM-Poisson model can handle under-, equi- and over-dispersed datasets with different mean values, although the credible intervals are found to be wider for low sample mean values. The computational burden of its implementation is also not prohibitive. Using intersection crash data collected in Toronto and segment crash data collected in Texas, the results show that COM-Poisson models perform as well as Poisson-gamma models in terms of goodness-of-fit statistics and predictive performance. With the use of crash data collected at railway-highway crossings in South Korea, several COM-Poisson models were estimated and it was found that the COM-Poisson model can handle crash data when the modeling output shows signs of under-dispersion. The results also show that the COM-Poisson model provides better statistical performance than the gamma probability and traditional Poisson models. Furthermore, it was found that the COM-Poisson model has limitations similar to that of the Poisson-gamma model when handling data with low sample mean and small sample size. Despite its limitations for low sample mean values for over-dispersed datasets, the COM-Poisson is still a flexible method for analyzing crash data.

DEDICATION

Dedicated to my bother-in-law, Late Mr. D. Srinivas Reddy

ACKNOWLEDGEMENTS

First of all, I would like to extend deep and sincere gratitude to my advisor, Dr. Dominique Lord, for his many helpful suggestions, important advice, constant encouragement and guidance from the very early stage of this research. He has always been inspiring and his doors are always open for my questions and fruitful discussions. Without his encouragement and guidance, this dissertation would not have been written or completed.

My appreciation is extended to Dr. Seth Guikema for his encouragement and help in many parts of this research and for providing excellent resources for completing all the simulations.

I gratefully thank Dr. Jeff Hart and Dr. Samiran Sinha for their advice at various stages of this research. I am thankful that in the midst of all their activity, they agreed to be members of the dissertation committee.

I am also grateful to the Texas Transportation Institute (TTI) for providing funding, accommodation and facilities for the successful carrying out of my doctoral studies.

My special thanks to Soma Sekhar Dhavala for helping me on many of the statistical issues of this dissertation.

I would also like to thank Sunil Patil for his help in improving this dissertation.

I am very thankful to my friends and classmates for their help, co-operation, support and friendship throughout my study.

Last, and above all, I am very grateful to my parents, sisters, brother and Ashwini for their encouragement during good and bad times; without their support none of this would have been even possible.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

The primary objective of this research is to introduce a new statistical model for analyzing traffic crashes - a model that develops a relationship between traffic crashes and factors associated with their occurrence such as traffic flow, geometric design of road sections, horizontal curvature, vertical grade, lane width, and shoulder width among others (Abdel-Aty et al., 2004; Maycock and Hall, 1984; Miaou, 1994, Maher and Summersgill, 1996; Poch and Mannering, 1996; Miaou and Lord, 2003; Lord et al., 2005a; Lord and Bonneson, 2007). Statistical models are used to explain the process from which the crash data are extracted, screen variables, identify hazardous sites and for predictive capabilities. These models are further used in studying the factors that cause traffic crashes in order to implement potential counter-measures and to reduce the number of crashes and their severities. These statistical models are often referred to as "Accident Prediction models (APMs)". Although models are developed after crashes occurred on the highway entities, the primary goal is to predict the crashes on newly built roads or on the roads which are to be upgraded so that the hazardous sites are properly identified and corrected before they are being used (Lord, 2000).

Traffic safety plays an integral role in a sustainable transportation development strategy. Although transportation needs increase with the development of a nation, the main negative impact of modern road transportation systems is injury and loss of life as a result of road accidents. Out of all the transportation systems, the road transportation or highway network is the most complex and dangerous system. According to the report issued by World Health Organization (2004), an estimated 1.2 million people are killed in

This dissertation follows the style of Accident Analysis and Prevention.

road crashes each year and as many as 50 million are injured worldwide. The projections, according to the report, indicate that these figures will increase by about 65% over the next 20 years if the proper measures are not implemented. This, in turn, makes road traffic injuries to be the third leading contributor to the global burden of disease and unintentional injury by 2020.

Traffic crashes cause significant economic and social costs. The economic cost of road crashes and injuries is estimated to be 1% of gross national product (GNP) in low-income countries, 1.5% in middle-income countries and 2% in high-income countries (World Health Organization, 2004). The global cost is estimated to be US$ 518 billion per year. Low-income and middle-income countries account for US$ 65 billion (World Health Organization, 2004). These costs are mainly due to the huge monetary and non-monetary costs caused by traffic accidents (Elvik, 2000). Monetary (or direct) costs relate to expenses for lost productivity due to disabilities and death, medical treatments, costs for repairing or replacing damaged vehicles, emergency and administrative service costs and expenditures on safety programs and equipment to reduce crash damages. Non-monetary (or indirect) costs can include pain, grief and suffering due to crash injuries and deaths (Miranda-Moreno, 2006). Thus, the traffic safety is one of the main priorities for many government agencies, private organizations and the society as a whole.

Human error by road users (such as vehicle drivers and pedestrians) is the primary reason for the occurrence of motor vehicle crashes (Salmon et al., 2005). The application of proper traffic control devices, and good roadway design features can often reduce the probability and the severity of a crash occurrence. One can better understand the effect of these devices only by extensively analyzing crashes at the location where the measures are implemented. The success of traffic safety and highway improvement programs can be determined by the analysis of accurate and reliable traffic accident data in a systematic and scientific approach. The development of statistical models using traffic crash data is one of the most important among the different approaches.

This chapter is organized as follows. The first section gives the problem statement. The second section presents a brief discussion of the importance of statistical models. Third section gives the objectives of this research, which is then followed by the outline of the dissertation.

1.1 Problem Statement

The traditional Poisson and Poisson-gamma (or Negative Binomial or NB) models are the most common probabilistic structure of the models used by transportation safety analysts for modeling motor vehicle crashes. In fact, Poisson-gamma models are usually preferred over Poisson regression models since crash data have often been shown to exhibit over-dispersion (see Lord et al., 2005b), meaning that the variance is greater than the mean.

Although the Poisson and NB regression models possess desirable distributional properties to describe motor vehicle accidents, these models are not without limitations (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). The other important limitation associated with NB models is the mis-specification of the probability density function (PDF) when the data exhibits under-dispersion, the condition in which the mean is greater than the variance.

To overcome some difficulties described above, researchers have proposed the use of zero-inflated models (Lord et al., 2005b; Shankar et al., 1997 and 2003; Qin et al., 2004) and gamma probability models (Oh et al., 2006). It is important to note that these models work as a dual-state process, one of which is characterized by having a long-term mean equal to zero (Lord et al., 2005b; Warton, 2005; Wedagama et al., 2006; Kadane et al., 2006). A dual-state process may not be appropriate for analyzing crash data.

Recently, many new methods have been introduced in traffic safety research, such as the Beta-binomial model (De Lapparent, 2005; Tong and Lord, 2007), neural and Bayesian neural network models (Abdelwahab and Abdel-Aty, 2002; Xie et al., 2008), latent class models (Depaire et al., 2008), and Support Vector Machine (SVM) models (Li et al., 2007). But none of these methods proposed have been able to replace the traditional Poisson-gamma models for analyzing motor vehicle crashes. Given the limitations of the Poisson-gamma model, there is a need to evaluate whether alternative count data models could be used for modeling motor vehicle crashes.

1.2 Importance of Statistical Models

Statistical models play a vital role in traffic safety analysis. The aim of statistical models is to explain observed and random variations of accidents across sites based on the available information on traffic flows and site-specific attributes. They serve one or more of the three purposes given below:

1. Explaining the process: Developing a statistical model gives important information about the process or the system from which data are extracted. The covariates included in the model development will reflect the attributes of the "system".
2. Screening Variables: Another important application of developing statistical models is to know the specific or significant effects of the variables on the risk of the collision. Examining the sign or the significance of a variable is a vital step during the model development process and its application is exploratory in nature.
3. Prediction: The third application aims at developing statistical models for most predictive capabilities. These models could be used with data collected as part of the model development or with a completely new dataset.

Peltola (2000) describes a process which is solely dependent on injury crashes. Crash prediction models are of great use in that process. The process described in the study calculates the change in accidents and fatalities due to the implementation of a countermeasure. Initially, the accident history for the last five years is used for predicting the number of crashes. With the change in traffic and land use, but without measure, crashes predicted before are corrected. Later, the effect of the countermeasure is counted by applying a coefficient on the predicted number of crashes. Thus, (from figure 1.1) it is very clear that the statistical models have great importance in the road safety improvements.

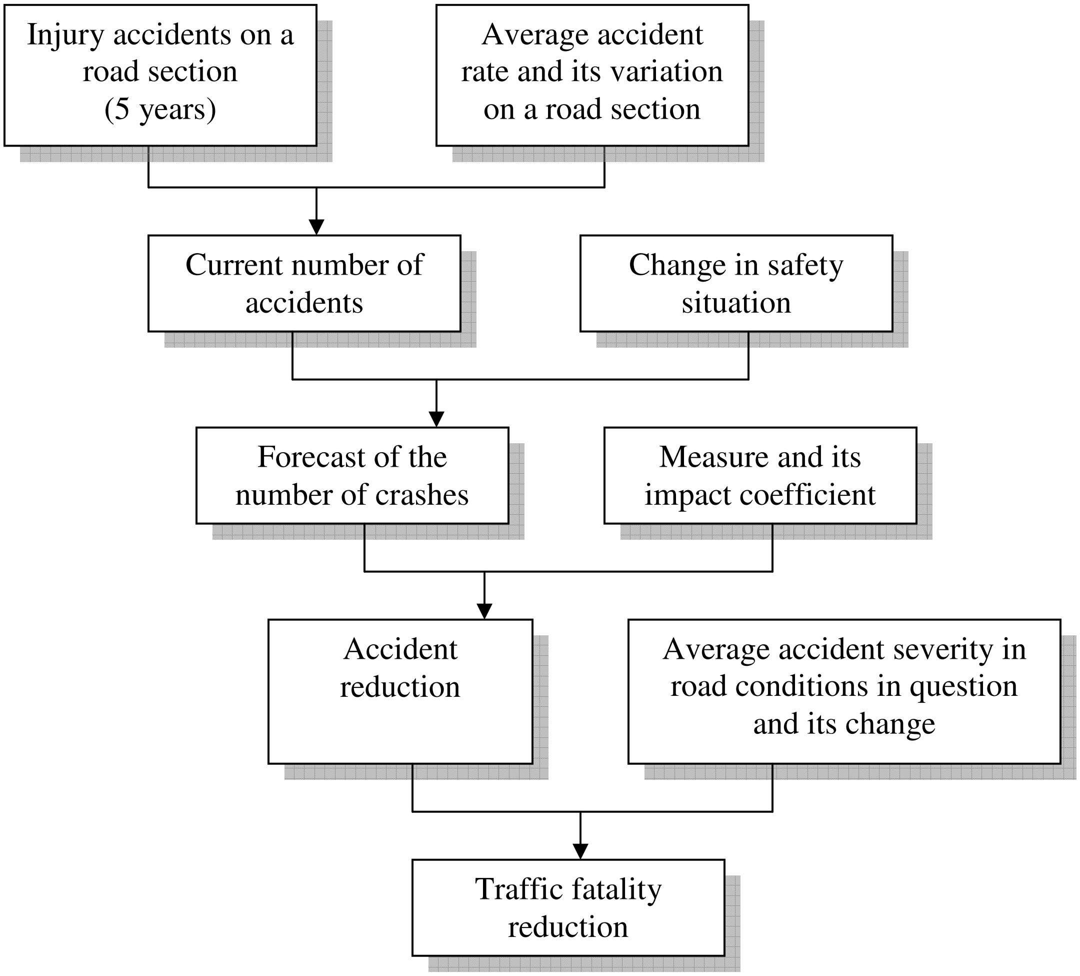

Injury accidents on a road section (5 years)
Average accident rate and its variation on a road section
Current number of accidents
Change in safety situation
Forecast of the number of crashes
Measure and its impact coefficient
Accident reduction
Average accident severity in road conditions in question and its change
Traffic fatality reduction

Figure 1.1: TARVA, organizational diagram for the modeling procedure (Peltola, 2000)

1.3 Research Objectives

The primary objective of this research is to evaluate the performance of the COM-Poisson distribution and describe the application of its generalized linear model (GLM) for analyzing motor vehicle crashes. In doing so, the following objectives are addressed in this research.

1. Assess the performance of Conway-Maxwell-Poisson distribution for datasets with different sample means and levels of dispersion. This analysis will be done in a Bayesian framework. This analysis will include testing datasets which are under-dispersed, equi-dispersed and over-dispersed. This objective primarily deals in (a) characterizing the parameter estimation accuracy of the Markov Chain Monte Carlo (MCMC) implementation of the COM-Poisson GLM, (b) estimating the computational burden of this MCMC implementation, and (c) investigating the degree of inaccuracy in using the asymptotic mean approximation proposed by Shmueli et al. (2005).

2. Examine the application of the COM-Poisson GLM for analyzing motor vehicle crashes. Compare the performance of the COM-Poisson models with the standard Poisson-gamma model, which is frequently used in analyzing traffic crash data, usually characterized by over-dispersion. The comparison analysis will be carried out using the most common functional forms employed by transportation safety analysts, which link crashes to the entering flows at intersections or on segments.

3. Evaluate the performance of the COM-Poisson GLM for analyzing the crash data exhibiting under-dispersion, in cases where Poisson and Poisson-gamma models cannot be used and then compare its performance with the Gamma probability models which have already been used for an under-dispersed crash dataset (Oh et al., 2006).

4. Evaluate the performance of the COM-Poisson distribution in terms of stability and presence of biasedness for data characterized by small sample size (SSS) and low mean problem (LMP). The bias in the estimation of model's coefficients will be investigated for these extreme conditions. The effect of using different prior distributions for the shape parameter will also be investigated.

5. Develop recommendations for implementing the COM-Poisson distribution in traffic safety research. Propose different directions for future research.

1.4 Dissertation Outline

The outline of this dissertation is as follows:

Chapter II gives a brief description about the different highway entities and functional forms used for modeling crash mean, which is followed by the description related to Poisson, negative binomial, zero-inflated, gamma probability and COM-Poisson models. This chapter also provides a discussion about limitations of some of those models and a brief summary of the methods for estimating the parameters.

Chapter III describes the performance of COM-Poisson GLM for simulated datasets for different mean values and levels of dispersions. The results concerning the parameter estimation accuracy of the Markov Chain Monte Carlo (MCMC) implementation of the COM-Poisson GLM, computational burden, and the degree of inaccuracy in using the asymptotic mean approximation are also presented.

Chapter IV investigates the performance of COM-Poisson for analyzing traffic crash data exhibiting over-dispersion. The results concerning the comparison of COM-Poisson with the NB models are also presented. This chapter further investigates the marginal effects of each parameter and gives the comparative analysis of the crash predictions with these two models.

Chapter V summarizes the results of the COM-Poisson GLM for analyzing the crash data exhibiting under-dispersion. The comparison results of COM-Poisson models with traditional Poisson and gamma probability models are also summarized. In doing so, the crash predictions with each of these models will be presented.

Chapter VI evaluates the performance of COM-Poisson for crash data characterized by low sample mean and small sample size in terms of stability and bias. It further investigates the effect of various prior distributions on the shape parameter and summarizes the results. The recommended sample size for a given sample mean for COM-Poisson will be proposed in this chapter.

Chapter VII gives a brief discussion and concluding remarks of this research. It also documents different directions for future work.

CHAPTER II

BACKGROUND

As discussed in Chapter I, the success of traffic safety and highway improvement programs can be determined only by the systematic analysis of crash data. The statistical models play vital role in the systematic analysis of traffic crash data.

This chapter is divided into eight sections. Section 2.1 lists the definition of transportation elements and the next section presents different functional forms used for modeling the crash mean. A brief description of the Poisson and NB models with their limitations are presented in section 2.3. An overview of the dual state models and few other statistical models that are used in traffic safety literature is presented in section 2.4 and 2.5 respectively. Section 2.6 presents a brief description of COM-Poisson distribution and its GLM framework. A brief note about the estimation methods is given in section 2.7. The last section summarizes the topics presented in this chapter.

2.1 Definition of Transportation Elements

Traffic safety analysts are primarily interested in analyzing the crashes occurring on the roadway network. Elements of the road network could include signalized intersections, unsignalized intersections, interstates, state highways, arterials and collectors that are located in urban and rural areas.

Intersections between highway segments and other primary roads (e.g., major and minor arterials, or major collectors) and where traffic volumes (Average daily traffic or ADT) are available on all approaches are usually referred as major intersections (Lord et al., 2008a). The intersections between the facility being analyzed and minor collectors, local roads, access driveways, or any intersection for which traffic volumes are not available on approaches intersecting the facility being analyzed are referred as minor intersections.

A homogenous section of a road which is delimited by major intersections or significant changes in the roadway cross-section (such as ADT), geometric characteristics (such as lane width, shoulder width, median presence and median width, side slope) of the facility, or the surrounding land uses is referred as a segment. The segments can be either undivided or divided depending on the presence of a median.

Crashes occurring within or near the intersection are referred as intersection crashes, and all other crashes are referred as segment crashes. Crashes that have already been defined as intersection or intersection-related in the accident report and that occurred within 250 ft (76 m) of the intersection center are assigned to the intersection (Lord et al., 2008a). Also the crashes that are not identified as intersection or intersection-related, but occurred within 250 ft from the middle of the intersection are also assigned to that intersection. Figure 1.2 gives a pictorial description of a segment and an intersection.

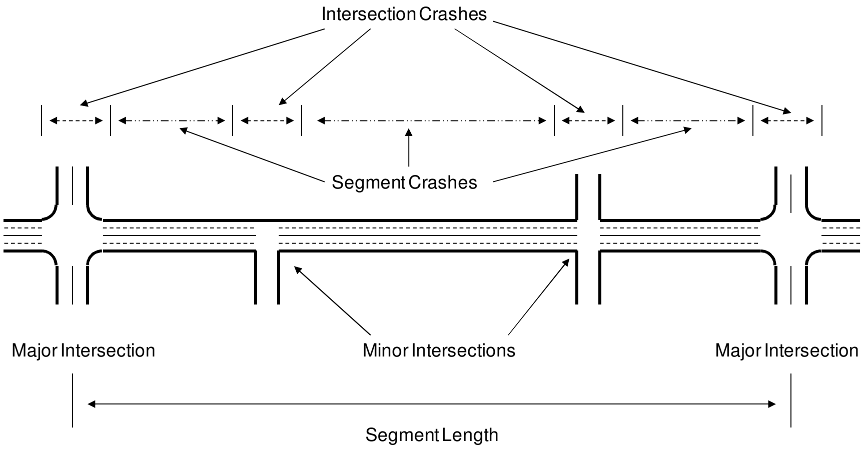


Figure 2.1: Definitions of intersections and segments (Lord et al., 2008a)

The input variables are different for analyzing intersection crashes and segment crashes. The segment crashes are classified as divided and undivided segment crashes and each has their own subset of variables. The crash analysis can also be done according to their severities (e.g. injury crashes), occurrence time (e.g. night time crashes) and causing factors (e.g. runoff road crashes) depending on the availability of the data. It may not be possible to include the traffic variations over time for any particular site because of prohibitive costs that are involved for collecting those data.

Input data variables for intersections are $\{y_i, F_{maj\_i}, F_{min\_i}, X_{ki}, t_i\}$

Input data variables for segments are $\{y_i, F_i, L_i, X_{ji}, t_i\}$

where,

y_i = the mean number of crashes per year for site i;

F_{Maj_i} = entering flow for the major approach (average annual daily traffic or AADT) for intersection i;

F_{Min_i} = entering flow for the minor approach (average annual daily traffic or AADT) for intersection i;

F_i = flow traveling on segment i (average annual daily traffic or AADT);

L_i = length in miles for segment i;

X_{ki} = a vector of intersection-specific covariates such as lighting condition, left/right turn lane etc.; $k = 1,\ldots\ldots\ldots\ldots,K$;

X_{ji} = a vector of segment-specific covariates such as median width, shoulder width etc.; $j = 1,\ldots\ldots\ldots\ldots..,J$;

$i = 1,\ldots\ldots\ldots\ldots,n$;

K, J = total number of covariates;

n = total number of sites; and

t_i - time period of observation for site i, $t > 0$

2.2 Functional Form for Modeling the Crash Mean

Several functional forms can be used for modeling the crash mean (λ_{it}) to capture the relationship between the crashes and the traffic flow. The functional form is different for analyzing intersection crashes and segment crashes. The typical flow-only functional forms that are used for analyzing intersection crashes are (Miaou and Lord, 2003):

$$\lambda_{it} = \beta_0 (F_{maj\_i} + F_{\min\_i})^{\beta_1} \tag{2.1}$$

$$\lambda_{it} = \beta_0 (F_{maj\_i})^{\beta_1} (F_{\min\_i})^{\beta_2} \tag{2.2}$$

$$\lambda_{it} = \beta_0 (F_{maj\_i} F_{\min\_i})^{\beta_1} \tag{2.3}$$

$$\lambda_{it} = \beta_0 (F_{maj\_i} + F_{\min\_i})^{\beta_1} \left[\frac{F_{\min\_i}}{F_{\min\_i}} \right]^{\beta_2} \tag{2.4}$$

$$\lambda_{it} = \beta_0 (F_{maj\_i})^{\beta_1} (F_{\min\_i})^{\beta_2} \exp(\beta_3 F_{\min\_i}) \tag{2.5}$$

Out of the five functional forms mentioned above, the second functional form is the most common and extensively used. It follows the logic of ‘“no traffic flows, no crashes,” and allows a non-linear relationship between crashes and traffic flows. Two advantages with this functional form are (1) the logic of having proper “boundary values,” (when the flow at both approaches is close to zero or when the flow at minor approach is close to zero) (2) not a logical one, but rather one that is based on previous experiences working with different data sets using a combination of visual inspections and statistical tests (Miaou and Lord, 2003).

The most common functional form that incorporates the site specific covariates in addition to the traffic flow in intersection models is often defined as follows:

$$\lambda_{it} = \beta_0 (F_{maj\_i})^{\beta_1} (F_{\min\_i})^{\beta_2} \exp(\beta_3 X_{3i} + ...\beta_n X_{ni}) \tag{2.6}$$

For analyzing segment crashes, the segment length needs to be included in the regression model. There has been much discussion about the effect of segment length on crashes and it is usually believed that segment length has a linear effect on the crashes (Lord and Bonneson, 2007; Fitzpatrick et al., 2008). Thus the segment length is often included as an offset rather than as a covariate. The commonly used functional form for segment crashes without covariates (general average annual daily traffic (AADT) model) is given as:

$$\lambda_{it} = \beta_0 L_i F_i^{\beta_1} \tag{2.7}$$

When the covariates are included in the segment model, then the following functional form is often defined as follows:

$$\lambda_{it} = \beta_0 L_i F_i^{\beta_1} \exp(\beta_2 X_{2i} + ...\beta_n X_{ni}) \tag{2.8}$$

2.3 Overview of Poisson and NB Models

This section gives a brief overview of Poisson and NB models followed by the limitations associated with these models.

2.3.1 Poisson model

Crashes are rare, discrete and independent events. The crash data are best characterized as Bernoulli trials with unequal crash probabilities that vary across drivers, vehicles, roadways, and environmental conditions (Lord et al., 2005b). Because of the very low probability of a crash and the large number of trials, these Bernoulli trials can be well approximated as Poisson trials. Thus the traditional Poisson distribution is considered to be the starting point for analyzing traffic crash data. The structure of the Poisson models is given as:

The number of crashes 'Y_i' for a particular i^{th} site when conditioned on its mean λ_i is assumed to be Poisson distributed and independent over all sites and time periods

$$Y_{it} | \lambda_{it} \sim Po(\lambda_{it}) \tag{2.9}$$

The mean number of the crashes λ_{it} is commonly specified as the exponential function of the covariates as

$$\lambda_{it} = f(X;\beta) \tag{2.10}$$

where,

$\beta = \beta_0 \ldots\ldots \beta_k$ are the vector of regression coefficients, and

X's are the vector of traffic flow and site specific covariates,

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

$$f(y_{it};\lambda_{it}) = \frac{e^{-\lambda_{it}}\lambda_{it}^{\ y_{it}}}{y_{it}!} \tag{2.11}$$

The mean and variance of the Poisson distribution is given by

$$E(y_{it}) = \lambda_{it} \tag{2.12}$$

$$Var(y_{it}) = \lambda_{it} \tag{2.13}$$

The important restriction with the traditional Poisson model is the equality of crash mean and crash variance as given above in Equations (2.12) and (2.13). This restriction is often violated by the crash data because of the existence of over-dispersion (and sometimes under-dispersion). Although, fitting the Poisson distribution to such data will not significantly influence the mean of regression coefficients, it will have a significant effect on the standard errors of the coefficients. The Poisson distribution underestimates the standard errors and in turn produces inflated t-values and confidence intervals of the coefficients when the data exhibit over- or under-dispersion (Miranda-Moreno, 2006).

The primary reason for observing over-dispersion is that the available covariates do not account for the full amount of individual heterogeneity. Also this 'extra' variation is thought to arise from unobserved differences across sites (Washington et al., 2003) and by some unmeasured uncertainties associated with the unobserved or unobservable variables, resulting in the omitted variable problem (Lord and Park, 2007). Furthermore, the over-dispersion is usually observed due to implicit randomness or impressions such as inaccuracy of traffic volumes, lack of information on other relevant site attributes, unmeasured variations in weather conditions, visibility, driver behavior, etc (Hauer, 1997). Finally, over-dispersion can also be attributed to the more random process related to the Bernoulli trials with non-equal success probabilities (Lord et al., 2005b).

2.3.2 Negative binomial model

The Poisson-gamma (or NB) distribution is the most common distribution used in the traffic safety literature, despite its limitations documented in the next section. This distribution is preferred over other mixed-Poisson distributions since the gamma distribution is the conjugate prior for the Poisson distribution. Also, as stated in Hauer (1997), the NB model offers a simple way to accommodate the over-dispersion and the mathematics to manipulate the relationship between the mean and the variance structures is relatively simple. Also, the likelihood function of NB model is readily available in statistical software programs, such as SAS (SAS, 2002), R (Venables et al., 2005) and Genstat (Payne, 2000). NB models can also be easily estimated in WinBUGS (Spiegelhalter et al., 2003) using a Bayesian modeling framework. The Poisson-gamma model has the following model structure (Lord, 2006):

The number of crashes 'Y_{it}' for a particular i^{th} site and time period t when conditioned on its mean μ_{it} is Poisson distributed and independent over all sites and time periods

$$Y_{it} \mid \mu_{it} \sim Po(\mu_{it}) \qquad \text{i = 1, 2, …, n and t = 1, 2, …, T} \tag{2.14}$$

The mean of the crashes μ_{it} is structured as

$$\mu_{it} = \lambda_{it} \exp(\varepsilon_{it}) \tag{2.15}$$

It is usually assumed that the $\exp(\varepsilon_{it})$'s are independent and gamma distributed with a mean equal to 1 and a variance α for all i and t.

$$\exp(\varepsilon_{it}) \mid \alpha \sim gamma(1/\alpha, 1/\alpha) \tag{2.16}$$

$$\text{and } \alpha \sim gamma(a,b) \tag{2.17}$$

where λ_{it} is the function of covariates $f(X;\beta)$,

ε_{it} is the model error term independent of covariates,

α is the dispersion parameter,

a and b are the shape and scale parameters respectively which need to be specified.

With this characteristic, it can be shown that Y_{it}, conditional on λ_{it} and α, is distributed as a Poisson-gamma random variable with a mean λ_{it} and a variance $\lambda_{it} + \alpha\lambda_{it}{}^{2}$ respectively. Although a large number of variance functions exist for Poisson-gamma models, they are not used in highway safety analysis (The reader is referred to Cameron and Trivedi (1998) and Maher and Summersgill (1996) for a description of alternative variance functions). The probability density function (PDF) of the Poisson-gamma structure described above is given by the following equation:

$$f(y_{it};\alpha,\lambda_{it}) = \frac{\Gamma(y_{it}+\alpha^{-1})}{\Gamma(\alpha^{-1})y_{it}!}(\lambda_{it}\alpha)^{y_{it}}(1+\alpha\lambda_{it})^{-(y_{it}+\alpha^{-1})} \tag{2.18}$$

The mean and variance of the Poisson-gamma random variable is given by

$$E(y_{it}) = \lambda_{it} \tag{2.19}$$

$$Var(y_{it}) = \lambda_{it} + \alpha\lambda_{it}{}^{2} \tag{2.20}$$

Note that as $\alpha \rightarrow 0$, the crash variance equals the crash mean and this model converges to the standard Poisson regression model

The term α is usually defined as the "dispersion parameter" (note that in some published documents, the variable α has also been defined as the “over-dispersion parameter”) of the Poisson-gamma distribution. This term has traditionally been assumed to be fixed and a unique value is applied to the entire dataset in the study. As described above, the dispersion parameter plays an important role in safety analyses, including the computation of the weight factor for the Empirical Bayes (EB) method (Hauer, 1997; Lord and Park, 2007) and the estimation of confidence intervals around the gamma mean and the predicted values of models applied to a different dataset than the ones employed in the estimation process (Geedipally and Lord, 2008).

Hauer (2001) first raised the issue of varying dispersion and reported that the dispersion parameter of Poisson-gamma models should be dependent upon the length of a highway segment. On the other hand, Heydecker and Wu (2001) attempted to estimate varying dispersion parameters as a function of sites' covariates, such as AADT, lane and shoulder widths among others. They asserted that the Poisson-gamma model with a varying dispersion parameter can better represent the nature of the crash dataset than the traditional Poisson-gamma model with a fixed dispersion parameter. The approach proposed by Heydecker and Wu (2001) was also used by Lord et al. (2005a) for modeling the safety performance of freeways as a function of traffic flow characteristics. Miaou and Lord (2003) have also noted that the dispersion parameter can be dependent upon the entering flows of crash-flow predictive models, suggesting that the variance function has an unobserved structure. Mitra and Washington (2007) suggested that a model with mis-specified mean function will have the variance function dependent upon the covariates of the models and concluded that the varying dispersion parameter may not be needed when the functional form describing the mean function contains several covariates.

Recently, Miranda-Moreno et al. (2005) reported that Poisson-gamma models with a varying dispersion parameter performed better than traditional models for identifying hazardous sites. El-Basyouny and Sayed (2006), on the other hand, indicated that this type of model offered better a statistical fit, but did not improve the hazardous site identification process. Lord and Park (2007) supported this finding and noted that Poisson-gamma models with a varying dispersion parameter influenced the EB estimates for multi-year analyses, but not for the identification of hazardous sites.

Poisson-gamma models with a varying dispersion parameter use the same PDF shown in equation (2.18) and estimate the same number of crashes for each observation, like the traditional Poisson-gamma model. However, instead of estimating a fixed dispersion parameter, these models use a varying dispersion parameter that can be estimated using the following expression (Heydecker and Wu, 2001; Mitra and Washington, 2007; Smyth, 1989):

$$\alpha_{it} = \exp(Z_{it} \times \delta_t) \tag{2.21}$$

where,

Z_{it} = a vector of secondary covariates (not necessarily have to be the same as the covariates in estimating the mean function μ_{it}),

δ_t = regression coefficients corresponding to covariates Z_{it} .

With equation (2.21), the model can be used for estimating a different dispersion parameter according to the sites' attributes (i.e., covariates). If there are no significant secondary covariates for explaining the systematic dispersion structure, the dispersion parameters will only contain a fixed value (i.e., constant term), resulting in a traditional Poisson-gamma regression model.

2.3.3 Limitations of the Poisson and NB model

Several studies have documented important limitations associated with the Poisson and NB model. The primary issue when dealing with the crash data analysis is the problem associated with the SSS and the LSM biases. The problem related to small sample sizes can be attributed to expensive costs of collecting crash data and the variables related to their occurrence in the field (Lord and Bonneson, 2005). There is a significant amount of ongoing research on these topics in the traffic safety literature. Maher and Summersgill (1996) defined the issue as "Low Mean Problem (LMP)" for such datasets. The problem of LMP was first raised by Maycock and Hall (1984) and later on carried out by Fridstrøm et al. (1995). For instance, it has been shown that when the sample mean value becomes small, traditional methods used to assess the goodness-of-fit (GOF) of generalized linear models (GLMs) estimated using the maximum likelihood estimating (MLE) method (both for Poisson and NB) can be highly unreliable and provide a biased estimate of the fit (Maycock and Hall, 1984; Maher and Summergill, 1996; Wood, 2002; Lord, 2006). In fact, it was shown that the Pearson's X^2 and the scaled deviance G^2 are no longer χ^2 distributed when the data are characterized by low mean values (Maher and Summersgill, 1996; Agrawal and Lord, 2006). Maher and Summersgill (1996) proposed a new test statistic as $G^2 / E(G^2)$ for the GOF tests. Later, Wood (2002) showed that this new statistic still fails for low mean values and proposed a grouped G^2 test statistic to estimate the fit of the models for the data characterized by a low mean value. More recently, Ye et al. (2008) proposed a power divergence test statistic for the Poisson models with low mean, but concluded that the more complex grouped test statistic still performs better for Poisson-gamma models.

Using simulated data, Clark and Perry (1989) reported that the Method of Moments (MM) and Maximum Quasi-Likelihood method become biased when sample mean $(\mu)\leq 3.0$ and sample size $(n)<20$. In another study, Piegorsch (1990) reported that the MLE of NB model was slightly less accurate for small sample sizes than the Quasi-likelihood estimators. Dean (1994) reported that the MLE produced a biased estimate and influenced the standard errors of the coefficients of the models as the sample size decreased. Toft et al. (2006) showed that the MLE method did not provide a reliable estimate of the parameters for the extreme conditions. Using the other simulation study, it was found that the use of Gibbs sampler with vague proper priors can lead to inaccurate posterior estimates of NB models when the data are characterized with low or moderate sample size (Natarajan and McCulloch, 1998). NB models have also been shown to be unable to handle data with extremely low mean values, which often produces sample data with many zeros.

Recent studies have also shown that the inverse dispersion parameter of NB models, ϕ = $1/\alpha$, can be significantly mis-estimated when the sample size is small and the sample mean value is low. This characteristic was observed both for MLE (Clark and Perry, 1989; Piegorsch, 1990; Lord, 2006; Lloyd-Smith, 2007) and Bayesian (FB) NB models (Airoldi et al., 2006; Lord and Miranda-Moreno, 2007). Using a simulation study, Lord (2006) reported that the MM, MLE and Weighted Regression estimators for estimating the dispersion parameter is very likely to be mis-estimated when the data are characterized by extreme conditions. Saha and Paul (2005) proposed a Biased-Corrected Maximum Likelihood (BCML) estimator for estimating the dispersion parameter (α) of a NB regression model and reported that the BCML also showed a biased result for the extreme conditions. Later, Lord and Miranda-Moreno (2007) used a Bayesian approach and reported that the dispersion parameter of Poisson-gamma models will be significantly biased when the data are characterized by LSM and SSS. These results in turn negatively influence the EB estimates for hot spot identification (Hauer and Persaud, 1987) and also the prediction of confidence intervals for comparing the safety performance of different highway design alternatives (Agrawal and Lord, 2006; Geedipally and Lord, 2008).

Although very rare, there is a possibility for the traffic crash data to exhibit under-dispersion when they are used in a context of generalized linear model (Oh et al., 2006; Park and Lord, 2007). This phenomenon is less convenient to model (Oh et al., 2006). The NB GLM could theoretically handle under-dispersion, since the dispersion parameter can be negative ($Var(Y) = \mu + (-\alpha)\mu^2$). However, in this case, the mean of the Poisson is no longer gamma distributed because this latter distribution cannot have negative parameters (i.e. $gamma(1/\alpha, 1/\alpha)$). In addition, researchers who have worked on the characterization of the NB distribution and GLM have indicated that a negative dispersion parameter could lead to a mis-specification of the PDF (when $-1/(\text{max of counts}) < \alpha$) (Clark and Perry, 1989; Saha and Paul, 2005). In summary, the NB GLM cannot or has difficulties converging with the datasets exhibiting under-dispersion and datasets that contain intermingled over- and under-dispersed counts (for dual-link models only, since the dispersion characteristic is captured using the covariate-dependent dispersion parameter).

2.4 Dual-State Models

This section presents a brief discussion of zero-inflated models and gamma probability model with their limitations.

2.4.1 Zero-inflated models

Over-dispersion is often characterized by “excess zeros”, that is the number of zeroes exceeds what is commonly expected under a normal Poisson process (Lord et al., 2005b). For the data with preponderance of zeros, the traditional Poisson and NB models will produce biased estimates. Recently, researchers have proposed the use of “Zero -inflated” or “Zero altered” probability models (Lord et al., 2005b; Shankar et al, 1997 and 2003; Qin et al, 2004) to model motor vehicle crashes. This dual-state process involves a zero-count state and a non-zero state (e.g., Shankar et al, 1997 and 2003; Qin et al, 2004). The zero count state includes sites defined as with a “perfect” safe condition which has the probability of accident occurrence of zero or very low that always generate zero accidents. The non-zero state includes locations at which the accident occurrence follows a Poisson (or NB) distribution (this state also includes zero count sites). The most common Zero-altered probability processes used in the traffic safety literature are the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB). These zero-inflated models have great flexibility (in a statistical sense) in uncovering processes affecting accident frequencies on roadway sections observed with zero accidents and those with observed accident occurrences (Shankar et al, 1997).

For a Zero-Inflated Poisson (ZIP) process, let Y be the number of crashes that has occurred on a particular road segment and δ be the probability that the road section will have a zero crash state and $1-\delta$ be the probability that the crashes follow a Poisson distribution , then the probability density function (PDF) of ZIP models has the following structure:

$$P(Y) = \delta + (1-\delta)e^{-\lambda}; Y = 0 \qquad (2.22)$$

$$P(Y) = (1-\delta)\frac{e^{-\lambda}\lambda^{y}}{y!};Y \geq 0 \qquad (2.23)$$

Similarly for a Zero-Inflated Negative Binomial (ZINB) process, let Y be the number of crashes that hasoccurred on a particular road segment and δ be the probability that the road section will have a zero crash state and $(1-\delta)$ be the probability that the crashes follow a true negative binomial distribution. Then the PDF of ZINB models has the following structure:

$$P(Y) = \delta + (1-\delta)(1+\alpha\lambda)^{-\alpha^{-1}};Y = 0 \qquad (2.24)$$

$$P(Y) = (1-\delta)\left[\frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})y!}(\lambda\alpha)^{y}(1+\alpha\lambda)^{-(y+\alpha^{-1})}\right];Y \geq 0 \qquad (2.25)$$

where α corresponds to the dispersion parameter and λ corresponds to the mean of the site.

These models have been shown to be inappropriate for modeling crash data, since the crash data do not exhibit two distinct generating processes, one of which is characterized by having a long-term mean equal to zero (Lord et al., 2005b; Warton, 2005; Wedagama et al., 2006).

2.4.2 Gamma probability model

The gamma probability model can be used for analyzing under-dispersed and over-dispersed data. Oh et al. (2006) analyzed crashes occurring at railway-highway crossings using a gamma probability model where the data were suspected to have under-dispersion (which is uncommon with the traffic crash data). They found that the gamma probability model provided a good fit for the railway-highway crossing crash data. The gamma probability model for the count data is given by:

$$P(y_i = k) = Gamma(\alpha k, \lambda_i) - Gamma(\alpha k + \alpha, \lambda_i) \qquad (2.26)$$

where λ_i is the mean of the crashes at i^{th} site and is given by $\lambda_i = \exp(\beta X_i)$

$$Gamma(\alpha k, \lambda_i) = 1 \text{ if k=0;} \tag{2.27}$$

$$Gamma(\alpha k, \lambda_i) = \frac{1}{\Gamma(\alpha k)} \int_0^{\lambda_i} t^{\alpha k - 1} e^{-t} dt \quad \text{if k>0;} \tag{2.28}$$

where α is the dispersion parameter; for α >1, it is under-dispersion; for α <1, it is over-dispersion; for α =1, it is equi-dispersion which means that the gamma model reduces to a Poisson model.

Also, it is important to note that the gamma probability model works as a dual-state model similar to zero-inflated models. As noted in previous section, a dual-state process may not be appropriate for analyzing crash data.

2.5 Overview of Other Crash Prediction Methods

Recently, there are many other statistical models which have been introduced to analyze motor vehicle crashes, either to establish relationships or for predicting crashes. These models include Poisson-lognormal model (Lord and Miranda-Moreno, 2007; Park and Lord, 2007), Beta-binomial model (De Lapparent, 2005; Tong and Lord, 2007), neural and Bayesian neural network models (Abdelwahab and Abdel-Aty, 2002; Xie et al., 2008), latent class models (Depaire et al., 2008), and Support Vector Machine (SVM) models (Li et al., 2007) are among the few. The univariate modeling framework has commonly been used in developing crash prediction models over the last few decades, but more recently the multivariate modeling framework has been used in crash data analysis (Tunaru, 2002; Miaou and Song, 2005; Song et al., 2006; Park and Lord, 2007).

2.6 Conway-Maxwell-Poisson Model

Given the important limitations above with various distributions that were used to analyze crash data, there is a need to evaluate other alternative models. Among those, the Conway-Maxwell-Poisson (COM-Poisson) model is the one that is evaluated in this research. Since the Conway-Maxwell-Poisson generalized linear model (COM-Poisson GLM) has the ability to handle under-dispersed data, it might prove to be advantageous over the traditional Poisson-gamma models. The COM-Poisson distribution has been used in many studies such as analyzing word length (Shmueli et al., 2005), birth process models (Ridout and Besbeas, 2004), prediction of purchase timing and quantity decisions (Boatwright et al., 2003), quarterly sales of clothing (Shmueli et al., 2005), internet search engine visits (Telang et al., 2004), the timing of bid placement and extent of multiple bidding (Borle et al., 2006), modeling electric power system reliability (Guikema and Coffelt, 2008) and modeling motor vehicle crashes (Lord et al., 2008b). Only Guikema and Coffelt (2008) and Lord et al. (2008b) have used the COM-Poisson in a regression setting.

2.6.1 Distribution

The COM-Poisson distribution is a generalization of the Poisson distribution and was originally developed in 1962 (Conway and Maxwell, 1962) as a method for modeling both under-dispersed and over-dispersed count data with a single link. It was then "rediscovered" by Shmueli et al. (2005), where many of the properties of the distribution were also first derived. The COM-Poisson belongs to the exponential family as well as to the two-parameter power series family of distributions. This distribution introduces an extra parameter ν which governs the rate of decay of successive ratios of probabilities. It nests the usual Poisson ($\nu = 1$), geometric ($\nu = 0$) and Bernoulli ($\nu = \infty$) distributions. The COM-Poisson distribution allows for both thicker and thinner tails than those of the Poisson distribution (Boatwright et al., 2003; Shmueli et al., 2005). The conjugate priors for the parameters of the COM-Poisson distribution have also been derived (Kadane et al., 2006).

The COM-Poisson distribution is a two-parameter extension of Poisson distribution that generalizes some well-known distributions including the Poisson, Bernoulli, and geometric distributions (Shmueli et al., 2005). It also offers a more flexible alternative to distributions derived from these discrete distributions (i.e. the binomial and negative binomial distributions). The COM-Poisson distribution can handle both under-dispersion (variance less than the mean) and over-dispersion (variance greater than the mean). The PDF of the COM-Poisson for the discrete count Y is given by Equations (2.29) and (2.30).

$$P(Y=y)=\frac{1}{Z(\lambda,\nu)}\frac{\lambda^{y}}{(y!)^{\nu}} \tag{2.29}$$

$$Z(\lambda,\nu)=\sum_{n=0}^{\infty}\frac{\lambda^{n}}{(n!)^{\nu}} \tag{2.30}$$

where λ is a centering parameter that is related directly to the mean of the observations and ν is the shape parameter of the COM-Poisson distribution. The condition $\nu > 1$ corresponds to under-dispersed data, $\nu < 1$ to over-dispersed data, and $\nu = 1$ to equi-dispersed (Poisson) data. Several common PDFs are special cases of the COM-Poisson with the original formulation. Specifically, setting $\nu = 0$ yields the geometric distribution, $\lambda < 1$ and $\nu \to \infty$ yields the Bernoulli distribution in the limit, and $\nu = 1$ yields the Poisson distribution. This flexibility greatly expands the types of problems for which the COM-Poisson distribution can be used in modeling count data.

The asymptotic expressions for the mean and variance of the COM-Poisson derived by Shmueli et al. (2005) are given by Equations (2.31) and (2.32) below.

$$E[Y]=\frac{\partial \log Z}{\partial \log \lambda} \tag{2.31}$$

$$Var[Y]=\frac{\partial^{2} \log Z}{\partial \log^{2} \lambda} \tag{2.32}$$

The COM-Poisson distribution does not have closed-form expressions for its moments in terms of the parameters λ and v. However, the mean can be approximated through a few different approaches, including (*i*) using the mode, (*ii*) including only the first few terms of Z when v is large, (*iii*) bounding E[Y] when v is small, and (*iv*) using an asymptotic expression for Z in Equation (2.30). Shmueli et al. (2005) used the last approach to derive the approximation in Equation (2.33).

$$E\left[Y\right] \approx \lambda^{1/v} + \frac{1}{2v} - \frac{1}{2} \tag{2.33}$$

Using the same approximation for Z as in Shmueli et al. (2005), the variance can be approximated as:

$$Var\left[Y\right] \approx \frac{1}{v}\lambda^{1/v} \tag{2.34}$$

These approximations may not be accurate for ν>1 or $\lambda^{1/v} < 10$ (Shmueli et al. 2005).

Despite its flexibility and attractiveness, the COM-Poisson has limitations in its usefulness as a basis for a GLM, as documented in Guikema and Coffelt (2008). In particular, neither λ nor v provide a clear centering parameter. While λ is approximately the mean when v is close to one, it differs substantially from the mean for small v. Given that v would be expected to be small for over-dispersed data, this would make a COM-Poisson model based on the original COM-Poisson formulation difficult to interpret and use for over-dispersed data.

Guikema and Coffelt (2008) proposed a re-parameterization using a new parameter μ = $\lambda^{1/v}$ to provide a clear centering parameter. This new formulation of the COM-Poisson is summarized in Equations (2.35) and (2.36) below:

$$P(Y=y)=\frac{1}{S(\mu,\nu)}\left(\frac{\mu^y}{y!}\right)^\nu \tag{2.35}$$

$$S(\mu,\nu)=\sum_{n=0}^{\infty}\left(\frac{\mu^n}{n!}\right)^\nu \tag{2.36}$$

By substituting $\mu=\lambda^{1/\nu}$ in equations (4), (5), and (41) of Shmueli et al. (2005), the mean and variance of *Y* are given in terms of the new formulation as $E[Y]=\frac{1}{\nu}\frac{\partial \log S}{\partial \log \mu}$ and $V[Y]=\frac{1}{\nu^2}\frac{\partial^2 \log S}{\partial \log^2 \mu}$ with asymptotic approximations $E[Y]\approx \mu+1/2\nu-1/2$ and $Var[Y]\approx \mu/\nu$ especially accurate once *μ>10.* With this new parameterization, the integral part of μ is the mode leaving μ as a reasonable centering parameter. The substitution $\mu=\lambda^{1/\nu}$ also allows ν to keep its role as a shape parameter. That is, if ν < 1, the variance is greater than the mean while ν > 1 leads to under-dispersion.

This new formulation provides a good basis for developing a COM-Poisson GLM. The clear centering parameter provides a basis on which the centering link function can be built, allowing ease of interpretation across a wide range of values of the shape parameter. Furthermore, the shape parameter ν provides a basis for using a second link function to allow the amount of over-dispersion, equi-dispersion or under-dispersion to vary across measurements.

2.6.2 Generalized linear model

Guikema and Coffelt (2008) developed a COM-Poisson GLM framework for modeling discrete count data using the reformulation of the COM-Poisson given in equations (2.35) and (2.36). This dual-link GLM framework, in which both the mean and the variance depend on covariates, is given in equations (2.37-2.38), where *Y* is the count random variable being modeled, x_i and z_j are covariates. There are *p* covariates used in the centering link function and *q* covariates used in the shape link function. The sets of parameters used in the two link functions do not need to be identical. If a single-link model is desired, the second link given by equation (2.38) can be removed allowing a single ν to be estimated directly.

$$\ln(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{2.37}$$

$$\ln(\nu) = \alpha_0 + \sum_{j=1}^{q} \alpha_j z_j \tag{2.38}$$

The GLM described above is highly flexible and readily interpreted. It can model under-dispersed datasets, over-dispersed datasets, and datasets that contain intermingled under-dispersed and over-dispersed counts (for dual-link COM-Poisson models only). The variance is allowed to depend on the covariate values, which can be important if high (or low) values of some covariates tend to be variance-decreasing while high (or low) values of other covariates tend to be variance-increasing. The parameters have a direct link to either the mean or the variance, providing insight into the behavior and driving factors in the problem, and the mean and variance of the predicted counts are readily approximated based on the covariate values and regression parameter estimates.

Parameter estimation in the COM-Poisson GLM presented above is challenging. The likelihood equation for the COM-Poisson GLM is complex, making analytical and numerical maximum likelihood estimation difficult. By the time this dissertation was written, Sellers and Shmueli (2008) developed the code for maximum likelihood estimation. The Bayesian estimation provides an attractive alternative for estimating the coefficients of the model. Guikema and Coffelt (2008) implemented the COM-Poisson GLM in WinBUGS using a custom-coded COM-Poisson distribution whereas the other coding is available for COM-Poisson GLM implemented in MATLAB® 7.1.0 R14 (The Mathworks Inc , Natick, MA).

2.7 Estimation Methods

This section gives a brief description about the maximum likelihood estimation and Bayesian estimation.

2.7.1 Maximum likelihood estimation (MLE) method

The maximum likelihood estimation approach is the most popular technique and has traditionally been used by many researchers for estimating the model coefficients.

The following steps give the procedure for MLE:
If $Y_1,....,Y_k$ are an iid sample from a population with PDF $f(y|\beta_1,...,\beta_k)$, the likelihood function is defined by

$$L(\beta|y) = L(\beta_1,.....,\beta_k|y_1,....y_n) = \prod_{i=1}^{n} f(y_i|\beta_1,.....,\beta_k). \qquad (2.39)$$

Definition of MLE: For each sample point y, let $\hat{\beta}(y)$ be a parameter value at which $L(\beta|y)$ attains its maximum as a function of β, with y held fixed. A maximum likelihood estimator (MLE) of the parameter based on a sample Y is $\hat{\beta}(y)$ (Casella and Berger, 2001).

If the likelihood function is differentiable (in β_i), the MLE of $\beta_1,...,\beta_k$ are obtained by maximizing $L(\beta|y)$. The MLE of $\beta_1,...,\beta_k$ are the simultaneous solutions of *k* equations such that

$$\frac{\partial}{\partial\beta_i} L(\beta|y) = 0, i = 1,....,k \qquad (2.40)$$

The desirable properties of maximum likelihood estimators are (Engineering statistics handbook, 2008):

1. they become unbiased minimum variance estimators as the sample size increases.

2. they have approximate normal distributions and approximate sample variances that can be calculated and used to generate confidence bounds.
3. likelihood functions can be used to test hypotheses about models and parameters.

2.7.2 Bayesian methods

Over the last five to ten years, researchers have shifted their interests to applying Bayesian methods in traffic safety (Miranda-Moreno et al, 2007; Miaou and Song, 2005; Lord and Miranda-Moreno, 2007). The two main approaches that are within the class of Bayesian methods are full Bayes (FB) and Empirical Bayes (EB) approach. The main difference between these two approaches is in the way the prior parameters are determined.

Full Bayes (FB) approach

The following steps give the procedure adopted for FB approach

Let $Y_1,....,Y_k$ are an iid sample from a population with PDF or PMF $f(y|\beta_1,...,\beta_k)$

In a Bayesian framework, a prior distribution on the parameter β_i is first assumed, denoted as π (β_i |η), where η is a vector of prior parameters. This prior information is then combined with the information brought by the sample into the posterior distribution, represented by p (β_i | y_i). The posterior distribution of β_i is given as (Carlin and Louis, 2000):

$$p(\beta_i \Big| y_i,\eta) = \frac{f(y_i|\beta_i)\pi(\beta_i|\eta)}{m(y_i|\eta)} \tag{2.41}$$

where $f(y_i|\beta_i)$ is the likelihood and the quantity $m(y_i|\eta)$ is the marginal distribution of y_i and is defined as

$$m(y_i|\eta) = \int f(y_i|\beta_i)\pi(\beta_i|\eta)\partial\beta_i \tag{2.42}$$

After the rediscovery of Markov Chain Monte Carlo (MCMC) methods by statistician, there has been an increased interest in applying hierarchical Bayes method for modeling motor vehicle crashes (Tunaru, 2002; Miaou and Song, 2005; Song et al, 2006; Miranda-Moreno et al, 2007; Qin et al, 2004; Miaou and Lord, 2003). The FB method is often preferred over the EB method because of its extensive flexibility in modeling traffic crashes (Miranda-Moreno, 2006). Not only the space-time variations are incorporated for modeling geographical and/or time dependence but also the consideration of randomness is given in FB models (Miranda-Moreno, 2006). The important issue related to the FB method is the specification of prior distributions (prior knowledge) in the modeling framework. The specification of the hyper-priors for the hyper-parameters has been discussed by many statisticians (see, e.g., Gelman et al, 2003; Rao, 2003). In crash data analysis, "vague" or "diffuse" hyper-priors are commonly recommended with the idea that these types of priors would reduce the influence of hyper-priors on the posterior distributions (Miranda-Moreno et al, 2007; Miaou and Song, 2005). Only recently did researchers showed interest in incorporating prior knowledge in Bayesian analysis (Lord and Miranda-Moreno, 2007). In a recent study, Miranda-Moreno et al. (2007) stated that there is a lack of methodologies for building informative hyper-priors as a potential solution to the LSM and SSS problems in the traffic safety literature. They proposed an informative prior and found that these priors perform well when compared to that of vague priors.

Empirical Bayes (EB) approach

EB method is not used in the parameter estimation process but to smooth the random fluctuation of crash counts and generate a more accurate estimate of the long-term mean at a given site. The prior parameters are estimated using maximum likelihood technique, weighted regression or method of moments involving the use of the accident data and the estimates depend only on the data. The EB method was initially introduced into the traffic safety literature by Hauer and Persaud (1984). Due to the less complicated construction, the EB method has been widely applied for identifying hotspot locations (Hauer and Persaud, 1987; Hauer, 2001; Persaud et al, 1999; Heydecker and Wu, 2001; Lord and Park, 2007). The EB method provides a better long-term estimate for a given site and reduces the regression-to-the mean bias (RTM). Also, the EB method is used to simplify the computational burden of the full bayes approach (the full bayes approach is described in the next part of this section). The main disadvantage noted with the EB method is related to its extensive data requirements (Hauer and Persaud, 1984; Persaud et al., 1999) and its assumption about estimating the dispersion parameter without uncertainty (Carlin and Louis, 2001). Also the EB approach has been advocated to perform poorly in the presence of data with low sample mean and small sample size (Lord 2006).

2.8 Summary

There are different definitions for defining the intersection and segment crashes. The definition followed in this research is that crashes that have already been defined as intersection or intersection-related in the accident report and that occurred within 250 ft (76 m) of the intersection center are assigned to the intersection. Crashes other than intersection-related and not occurred within 250 ft (76 m) of the intersection center are defined as segment crashes. There are different functional forms proposed for modeling the crash mean in the traffic safety literature. In the first and second section, a brief note about the definition of highway entities and the functional forms adopted for modeling the crash mean were presented. The most commonly used functional forms are given in equations (2.2) and (2.7) for developing general AADT models. For covariate models, the most common functional forms are given in equation (2.6) and (2.8) for intersection and segment models respectively.

A wide variety of statistical models have been used in the traffic safety analysis and Poisson-gamma (NB) is the most popularly used model for modeling motor vehicle crashes. Although the NB model possesses interesting desirable properties, it is not without limitations. A brief literature study on the limitations of NB model, a discussion about commonly used models and their limitations was presented in this chapter. A detailed discussion about the COM-Poisson distribution and its GLM framework was also given in the chapter. The last section gave a brief note and a short literature review on the estimation methods of the regression parameters. The following chapters give the results of this research.

CHAPTER III

PERFORMANCE OF THE COM-POISSON GLM

This chapter describes the performance of COM-Poisson regression models. The COM-Poisson distribution is known to handle both the under-dispersed and over-dispersed data. It is important to evaluate the performance of the COM-Poisson distribution for different situations such as high, moderate and low sample mean values which are frequently exhibited by crash datasets. All the computations in this chapter will be evaluated under the Bayesian setting with the inclusion of non-informative priors for the model parameters.

Guikema and Coffelt (2008) introduced a generalized linear model built on the COM-Poisson. The COM-Poisson GLM of Guikema and Coffelt (2008) is a full Bayesian model implemented in WinBUGS. This chapter evaluates the estimation accuracy and computational burden of the COM-Poisson GLM for datasets characterized by over-, under- and equi-dispersion with different means. It also characterizes the accuracy of the asymptotic approximation of the mean of the COM-Poisson suggested by Shmueli et al. (2005).

This chapter is organized as follows: First, the research methodology used in this chapter is presented. Second, the results of the computational study are given. Third, a brief discussion of the results is presented followed by a brief summary of the chapter.

3.1 Methodology

To test the estimation accuracy and computational burden of the MCMC implementation of the COM GLM of Guikema and Coffelt (2008), a number of datasets were simulated from the COM GLM with known regression parameters that correspond to a wide range of mean and variance values. The regression parameters of the COM GLM were estimated using the MCMC implementation. The estimated parameters were then compared to the known parameter values, and the computational burden of the MCMC was assessed in all the cases. In this section, details of the various steps involved were given.

3.1.1 Data simulation

In order to characterize the accuracy of the parameter estimates from the COM GLM, five different datasets were randomly generated for each of nine different scenarios. The nine scenarios include simulated datasets of under-dispersed, equi-dispersed and over-dispersed data. For each level of dispersion, three different sample means were used: high mean (~ 20.0), moderate mean (~ 5.0) and low mean (~ 0.8). Due to the high computational time and lack of readily available software, the analysis was restricted to five simulation runs (or datasets) for each scenario. Each of these five datasets was then used as input for the COM GLM, and the resulting parameters estimates were compared to the known parameters values that had be used to generate the datasets.

The 1,000 values of the covariates X_1 and X_2 were simulated from a uniform distribution on [0, 1]. The centering parameter λ and shape parameter ν were then generated according to Equations (2.37) and (2.38) with known (assigned) regression parameters. Note that the same covariates are used for both the centering and shape parameters. Realizations from the COM-Poisson are then generated using the inverse CDF method (Devroye, 1986).

The regression parameter values were selected in such a way that the shape parameter ν was always set between 0 and 1 for simulating the over-dispersed datasets, above 1 for the under-dispersed datasets and approximately 1 for the equi-dispersed datasets. The parameters that were assigned in simulating the datasets are given in the table below. Note that a single-link model was assumed by assigning values of zero to α_1 and α_2, respectively. However, these parameters were left in the MCMC COM GLM in order to test both (1) the computational burden of the full dual-link model and (2) the ability of the COM-Poisson model to accurately estimate zero values for these two parameters. Table 3.1 summarizes the characteristics of the simulation scenarios.

Table 3.1: Assigned parameters of the simulated datasets

	Over-dispersed data			Under-dispersed data			Equi-dispersed data		
	High mean	Moderate mean	Low mean	High mean	Moderate mean	Low mean	High mean	Moderate mean	Low mean
β_0	3.0	1.3	-2.0	3.0	1.7	0.2	3.0	1.7	0.2
β_1	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
β_2	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15
α_0	-0.4	-1.3	-1.3	1.0	1.0	1.2	0	0	0
α_1	0	0	0	0	0	0	0	0	0
α_2	0	0	0	0	0	0	0	0	0

All the simulations were carried in MATLAB® 7.1.0 R14 (The Mathworks Inc, Natick, MA)

3.1.2 Testing protocol

The MCMC implementation of the COM GLM proposed by Guikema and Coffelt (2008) was used for the model estimation process. The coefficients of the COM GLM were estimated using WinBUGS. Non-informative priors (i.e., N (0,100) priors) were utilized for the parameters of COM GLMs. A total of 3 Markov chains were used in the model estimation process with 50,000 iterations per chain and no thinning. The first 25,000 iterations (burn-in samples) were discarded. The remaining 25,000 iterations were used for estimating the coefficients. The Gelman-Rubin (G-R) convergence statistic was used to verify that the simulation runs converged properly. The convergence was known when the G-R statistic was below 1.1.

3.2 Results

This section first describes the selection of the error term used in approximating the normalizing term (S) in the COM-Poisson distribution (see Equation (2.36)). This is followed by an assessment of the performance of COM-Poisson for the nine scenarios mentioned above. The results concerning the accuracy of the asymptotic approximation of the mean of the COM-Poisson suggested by Shmueli et al. (2005) and the computational burden of COM-Poisson in WinBUGS are then discussed.

3.2.1 Error term

The S term in the COM-Poisson distribution is the sum of an infinite series. However, the contributions of new terms in the series decreases as more terms are added in the series. In order to approximate S, an iterative approach was used in which the change in S was monitored as new terms were added. The series was truncated when the contribution of new terms dropped below a predefined threshold, given as a fraction of the previous series value. This threshold is referred as the relative error ε. Four levels of error were considered in this study: ε= 0.1, 0.01, 0.001, and 0.0001. The first set of runs of the COM-Poisson MCMC was performed to determine the effect of ε on both the computational burden of the process and the parameter estimates. These runs were accomplished using an over-dispersed high mean dataset on a computer with a 1.50 GHz Pentium 4 CPU and 512 MB of RAM. As seen in Figure 3.1, the error term does not have much effect on the parameter estimation accuracy. The estimated parameters were almost the same at all error levels.

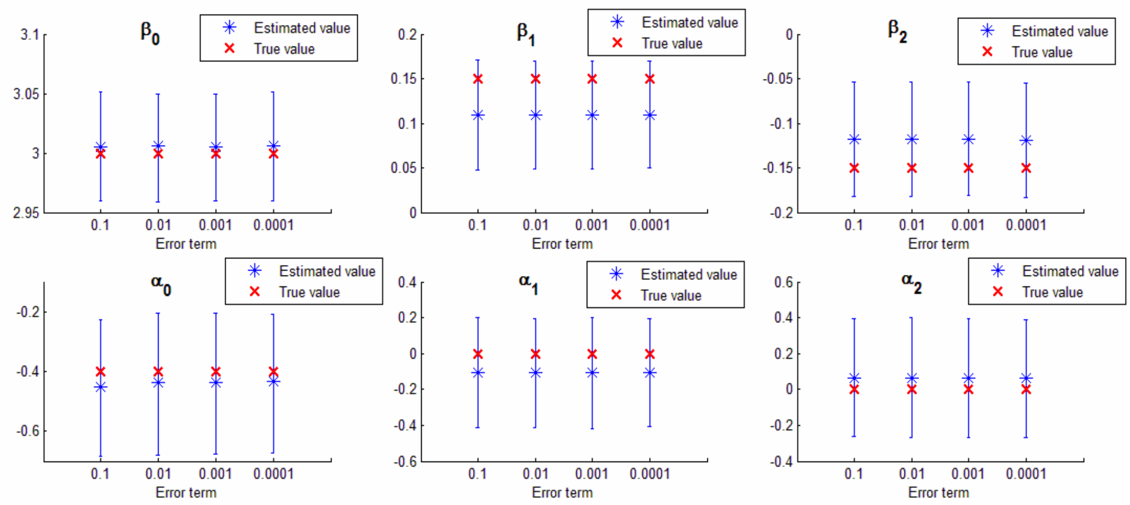


Figure 3.1: Parameter estimation accuracy versus relative error

The computational time was also not much different from one error level to another, as shown in Figure 3.2. Note that the computational time depends on the computer system configuration on which the simulations were performed. An error term of 0.01 was chosen in this analysis, although choosing a different error term value would not affect the results substantially.

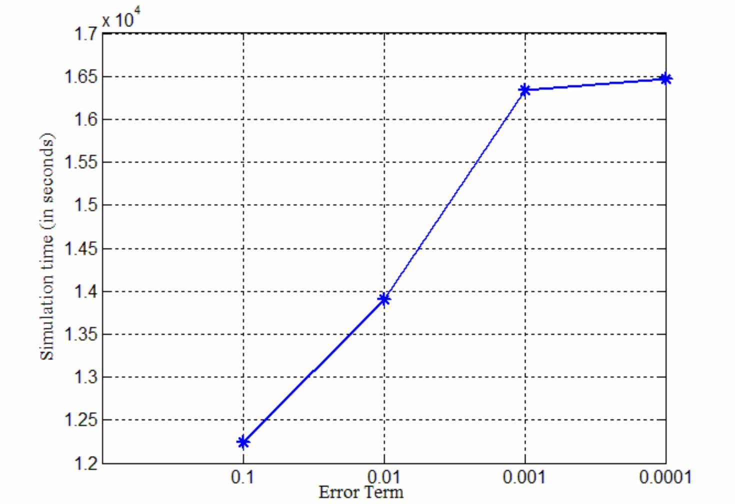


Figure 3.2: Simulation time for different error terms of COM-Poisson distribution

3.2.2 Parameter estimation accuracy

The estimated parameters and their 95% credible intervals were plotted and compared with the true parameters as shown in Figures 3.3-3.5. Each figure corresponds to a specific dispersion present in the data and each subplot corresponds to different sample mean value of the dataset. Figure 3.3 shows the plot for the over-dispersed datasets. It shows that the true parameters lie in the 95% credible interval for all cases and are generally close to the estimated posterior mean of the parameters. The credible intervals were found to be wider at low mean values for both the centering and shape parameter coefficients.

Figure 3.4 gives the plots for the under-dispersed datasets. Similar to the result above, the true parameters lie in the 95% credible interval for all cases and are generally close to the estimated posterior mean of the parameters. The credible intervals of the parameters were found to be wider (as expected) for low mean values for both centering and shape parameters.

Figure 3.5 shows the similar characteristics for the equi-dispersed datasets as that of other datasets. Except in one or two cases, all plots show that the true parameter lies inside the 95% credible intervals of estimated parameters. Although the problem with these exceptional cases was unknown yet, it could be attributed to the randomness in the datasets. Also, the credible intervals of the parameters were found to be wider for the low mean values for both centering and shape parameters.

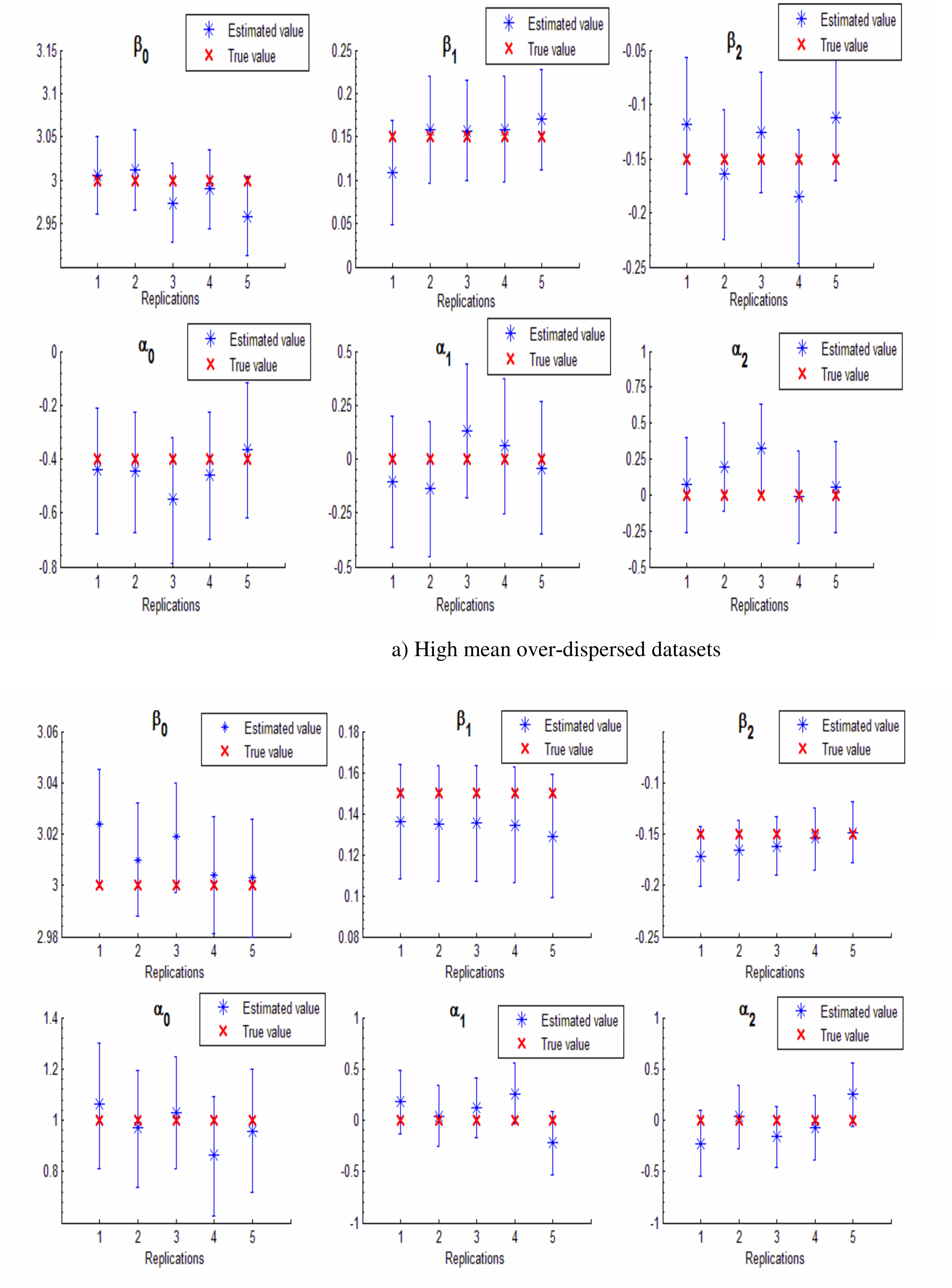
3.2.3 Bias of the parameters

The bias of an estimator is defined as the difference between an estimator's expected value and the true value of the parameter being estimated. If the bias is zero then the estimator is said to be unbiased. The bias of an estimator $\hat{\theta}$ is calculated as $E(\hat{\theta}) - \theta$ where θ is the true value of the parameter and the estimator $\hat{\theta}$ is a function of the observed data.

The bias of β and α parameters is calculated as the difference between their average values from 5 samples and the true (or assigned) value in each scenario.

The bias of centering parameter 'μ' is calculated as

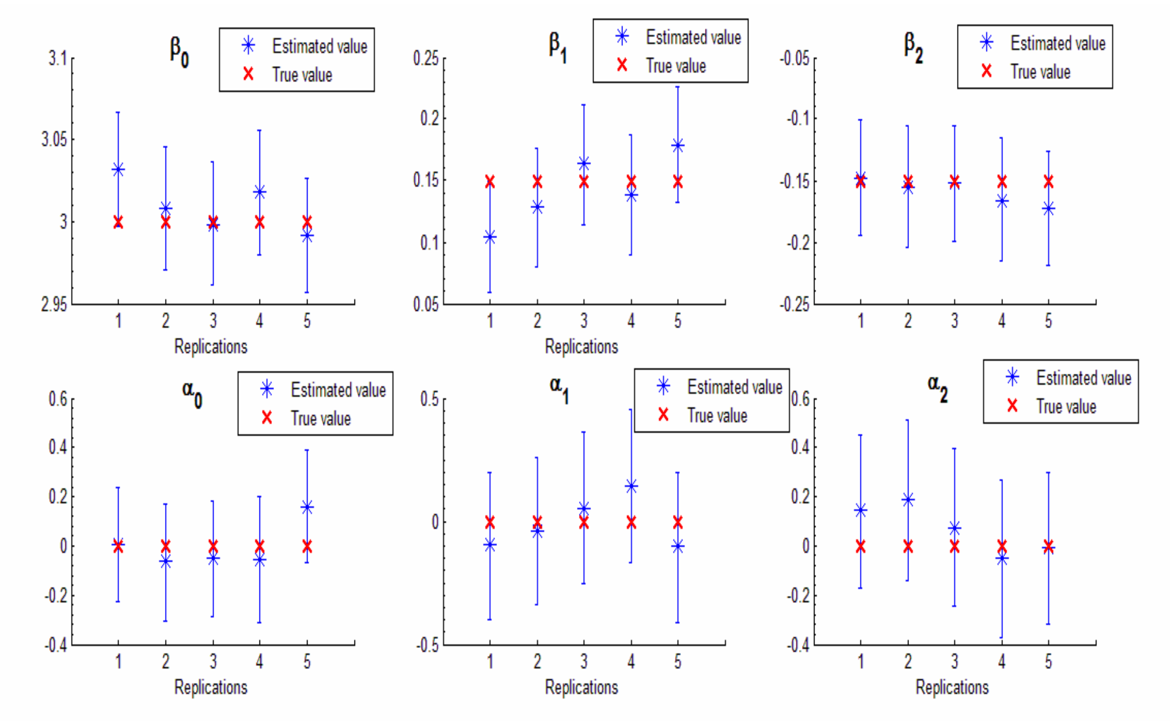
$$E(\hat{\mu}) - \mu = (\overline{\hat{\beta}}_0 - \beta_0) + (\overline{\hat{\beta}}_1 - \beta_1)\overline{X}_1 + (\overline{\hat{\beta}}_2 - \beta_2)\overline{X}_2 \qquad (3.1)$$



a) High mean over-dispersed datasets

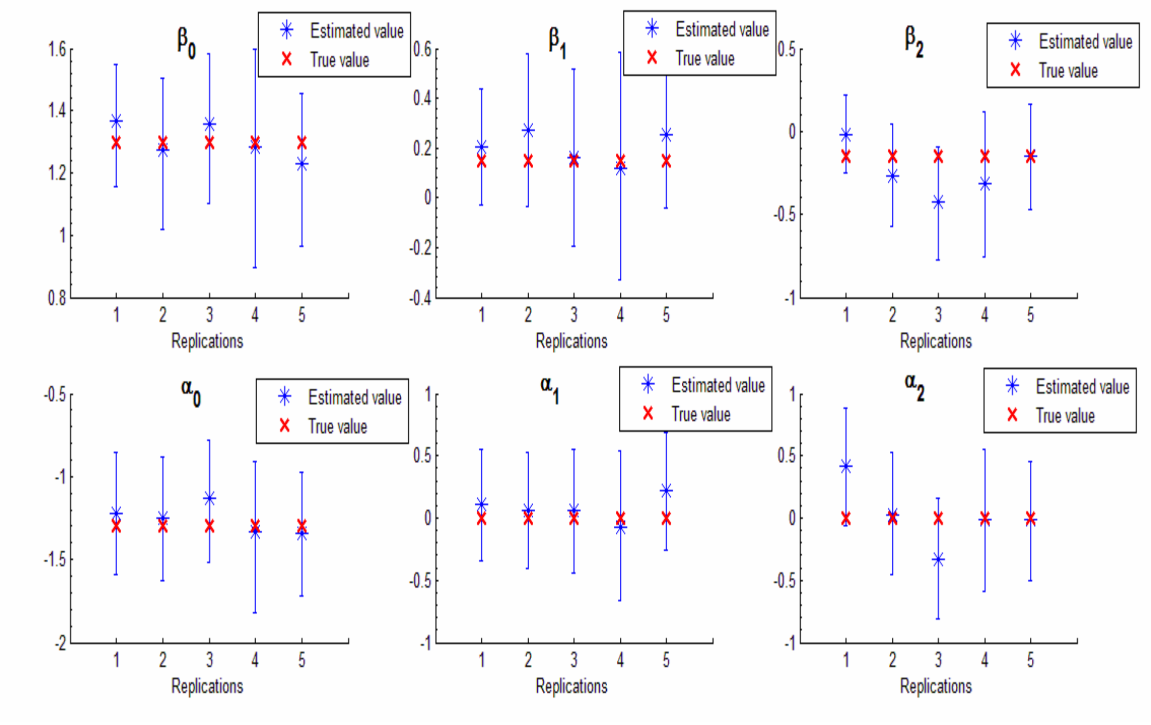
b) Moderate mean over-dispersed datasets

Figure 3.3: Parameter estimates of over-dispersed datasets



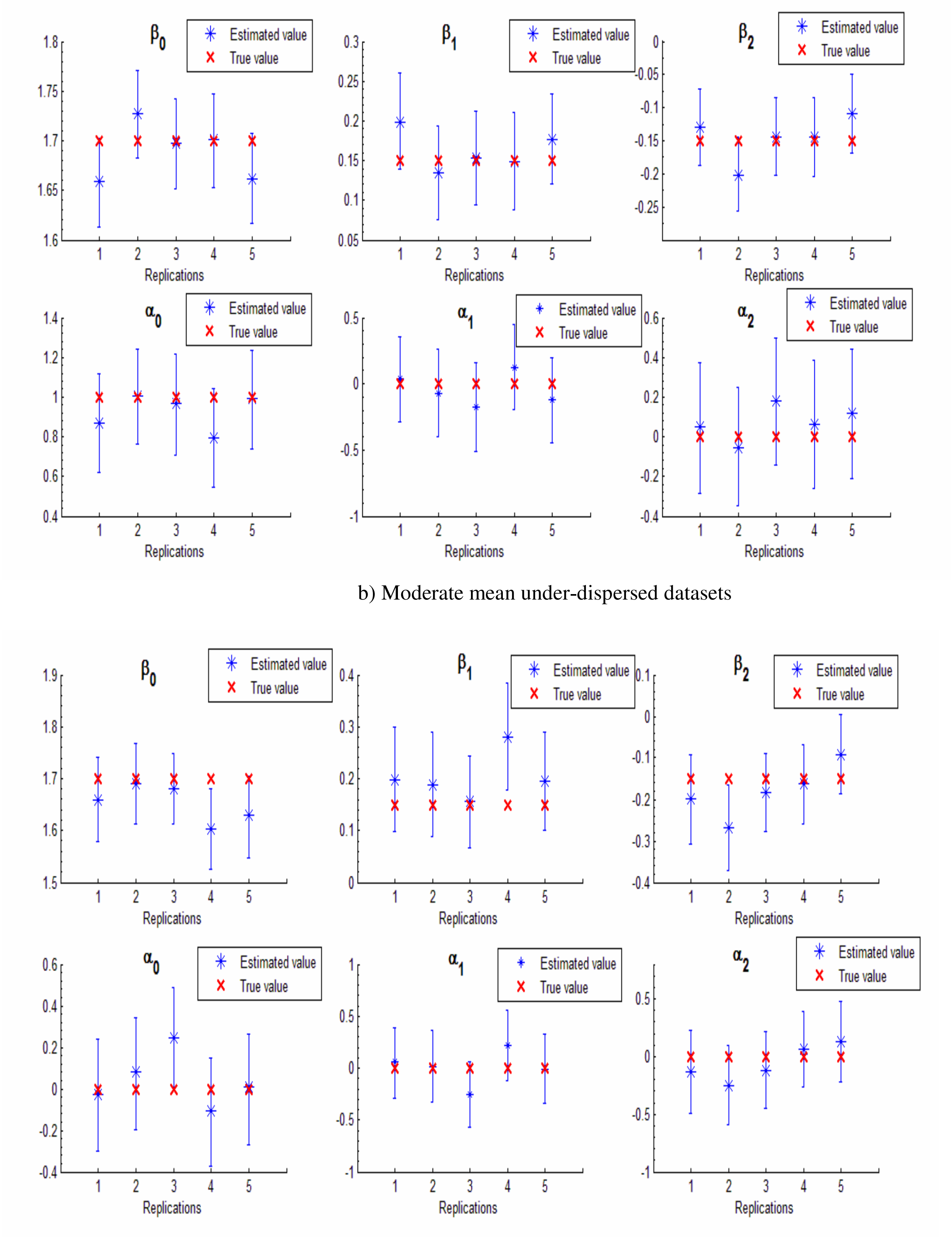
c) Low mean over-dispersed datasets

Figure 3.3: Continued



a) High mean under-dispersed datasets

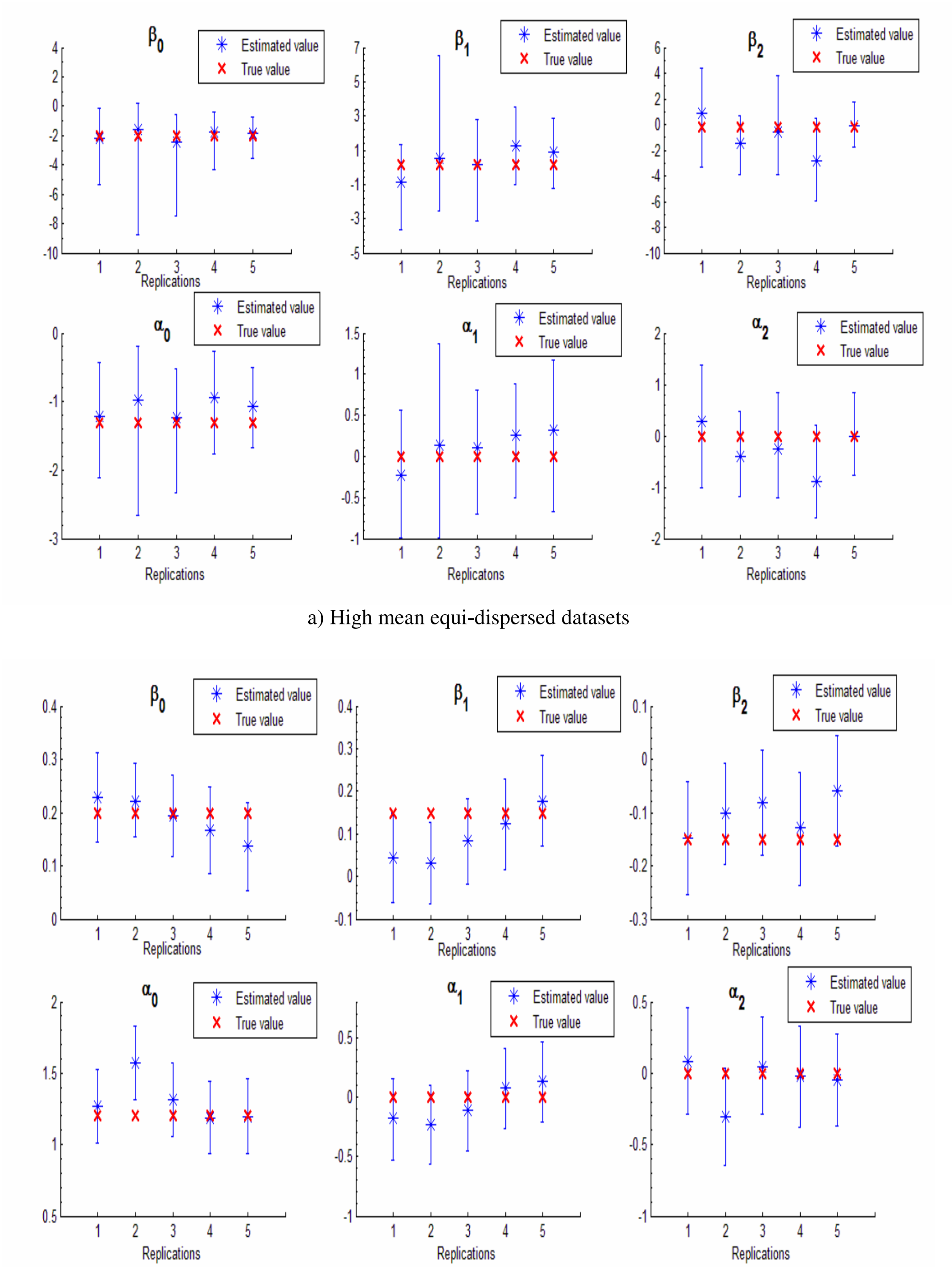
Figure 3.4: Parameter estimates of under-dispersed datasets



b) Moderate mean under-dispersed datasets

c) Low mean under-dispersed datasets

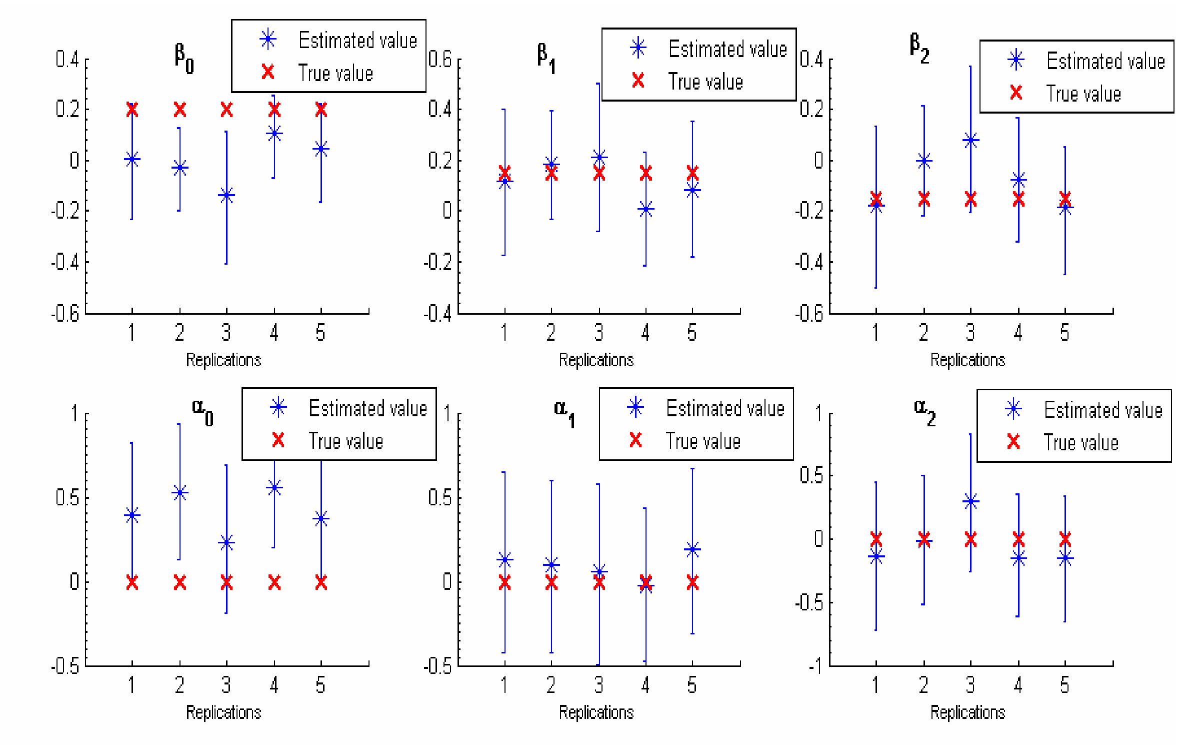
Figure 3.4: Continued



a) High mean equi-dispersed datasets

b) Moderate mean equi-dispersed datasets

Figure 3.5: Parameter estimates of equi-dispersed datasets



c) Low mean equi-dispersed datasets

Figure 3.5: Continued

The bias of centering parameter 'ν' is calculated as

$$E(\hat{\nu}) - \nu = (\overline{\hat{\alpha}_0} - \alpha_0) + (\overline{\hat{\alpha}_1} - \alpha_1)\overline{X}_1 + (\overline{\hat{\alpha}_2} - \alpha_2)\overline{X}_2 \qquad (3.2)$$

where $(\overline{\hat{\beta}_i} - \beta_i)$ and $(\overline{\hat{\alpha}_i} - \alpha_i)$ are the bias in the parameters and $\overline{X}_i$ is the average value of the independent variable, which is expected to be 0.5 since the independent variables are randomly simulated between 0 and 1. As seen from Figure 3.6, with the exception of the under-dispersed data, the bias increased as the mean values decreased. The bias did not change significantly for the under-dispersed data from one mean to other. The bias becomes worse for the over-dispersed datasets at low mean values.

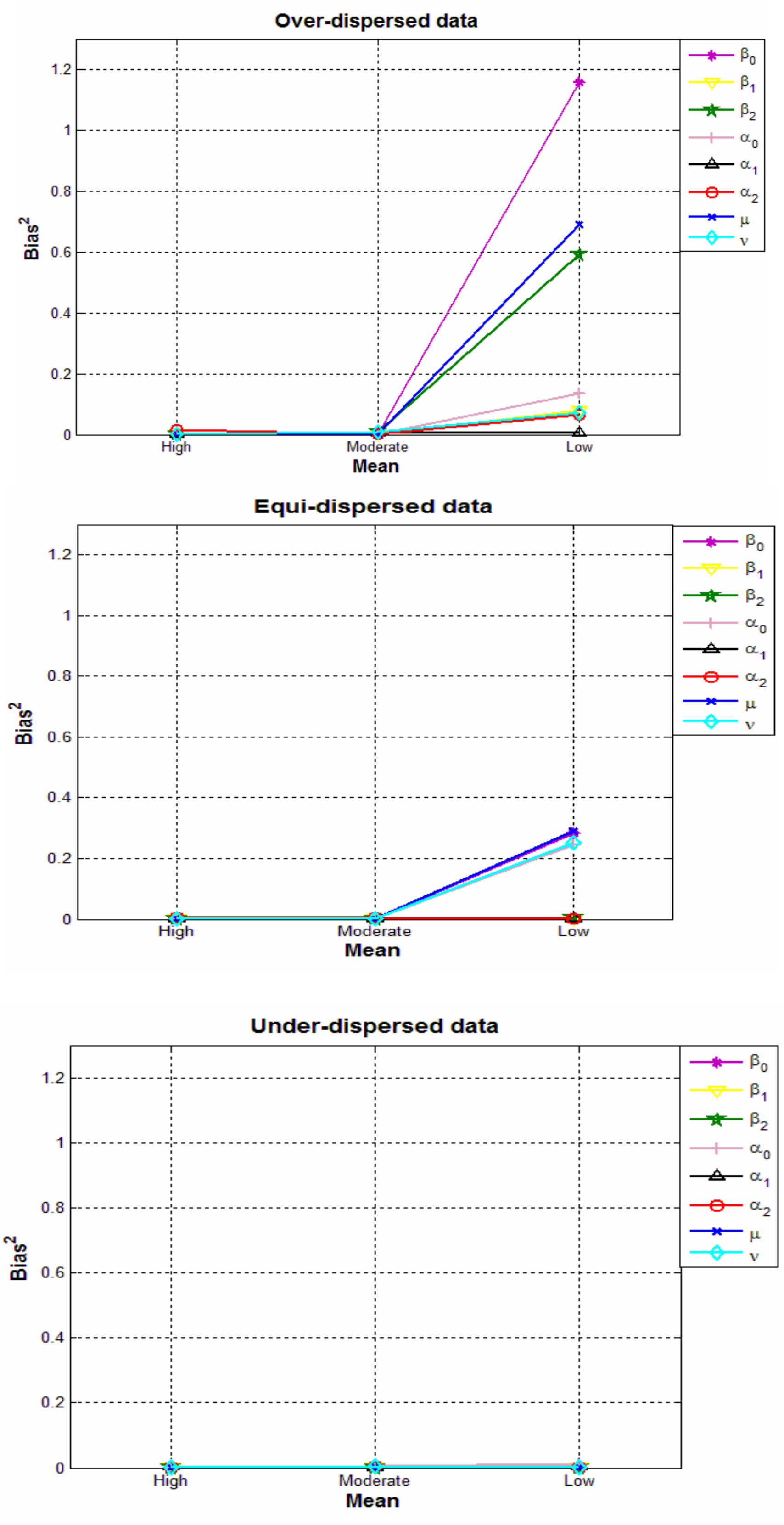


Figure 3.6: Bias^2 of the parameters

The biases in the mean values are plotted and are shown in Figures 3.7-3.9. Figure 3.7 gives plots of the estimated and true mean values for the over-dispersed datasets. For the true mean, first the true μ and v parameters were calculated from the true (or assigned) parameters. The 100,000 random counts were then simulated from the COM-Poisson distribution for the given μ and v. The mean of these random variables gives the true mean for each sample. Similarly, the predicted μ and v parameters were calculated from the estimated parameters for each of the five samples. Again, 100,000 random counts were simulated from the COM-Poisson distribution for the given μ and v. The mean of these random variables gives the predicted mean for each sample. The second subplot corresponds to the combined effect of all five samples. Instead of the parameters estimated for each sample, the average of the estimated parameter is considered in calculating the predicted mean in these plots.

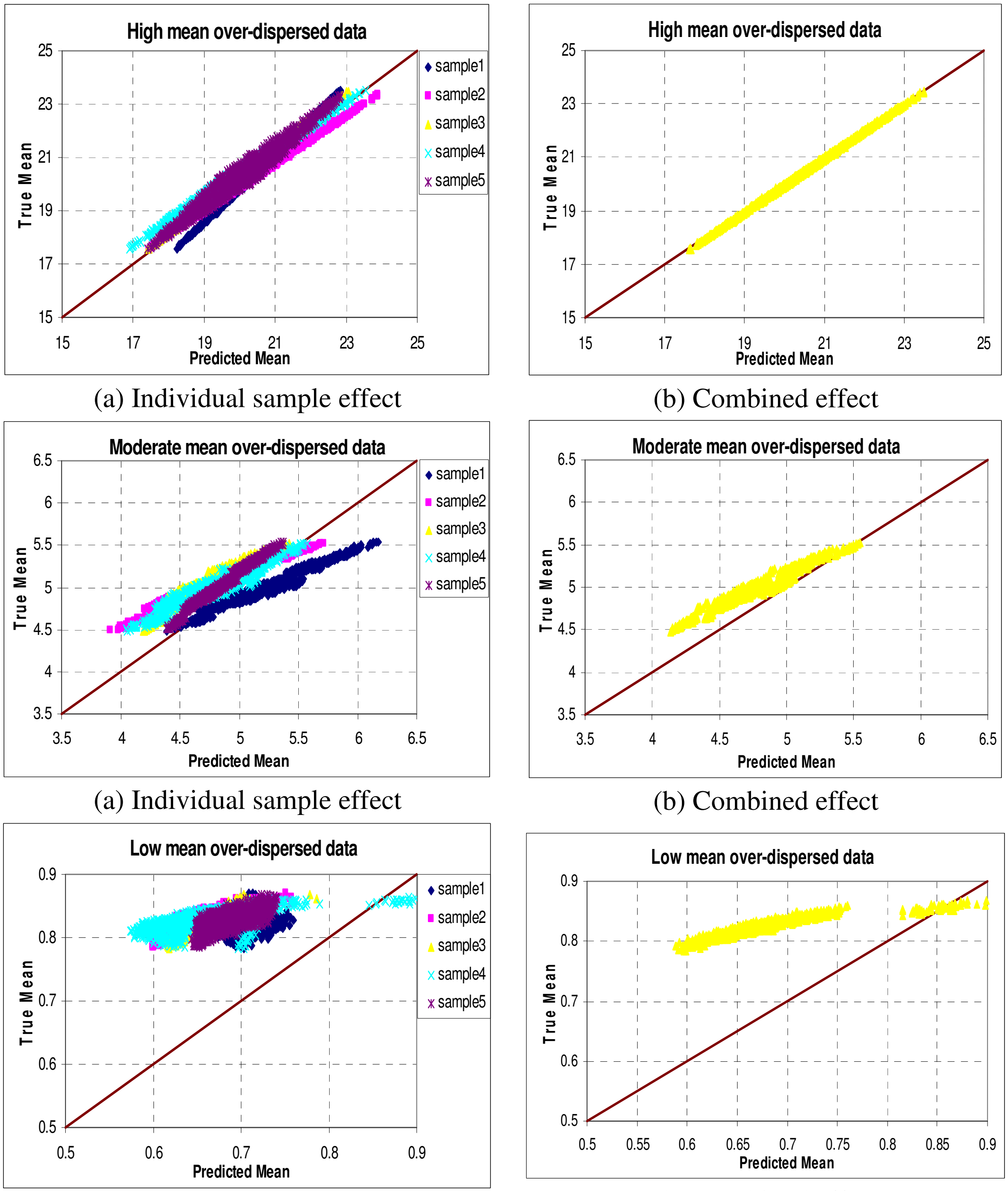
The COM-Poisson distribution performs better for high and moderate mean for all three categories of dispersion. For over-dispersed and equi-dispersed datasets, the performance is worse for all low sample mean values. The COM-Poisson distribution works well for all sample mean values for the under-dispersed datasets. However, the results in Chapter V about the application of COM GLM for analyzing motor vehicle crash data exhibiting under-dispersion (conditional on the mean) showed that the estimated mean is an unreliable estimate of traffic crashes at extremely low sample mean values (~0.3) for v >1. The centering parameter of the distribution was itself found to be a preferable estimate for predicting crashes.

3.2.4 Difference between true mean and predicted mean

This section gives the percentage difference between the true mean and predicted mean for all types of dispersions. The percent difference is calculated as:

Percent difference = (True mean-Predicted mean)*100 / True mean

Figure 3.10 gives the percent difference of the over-dispersed datasets. Similar to the results above, this figure shows that the performance of COM-Poisson distribution is good for high mean data and worse for low mean datasets.



(a) Individual sample effect

(b) Combined effect

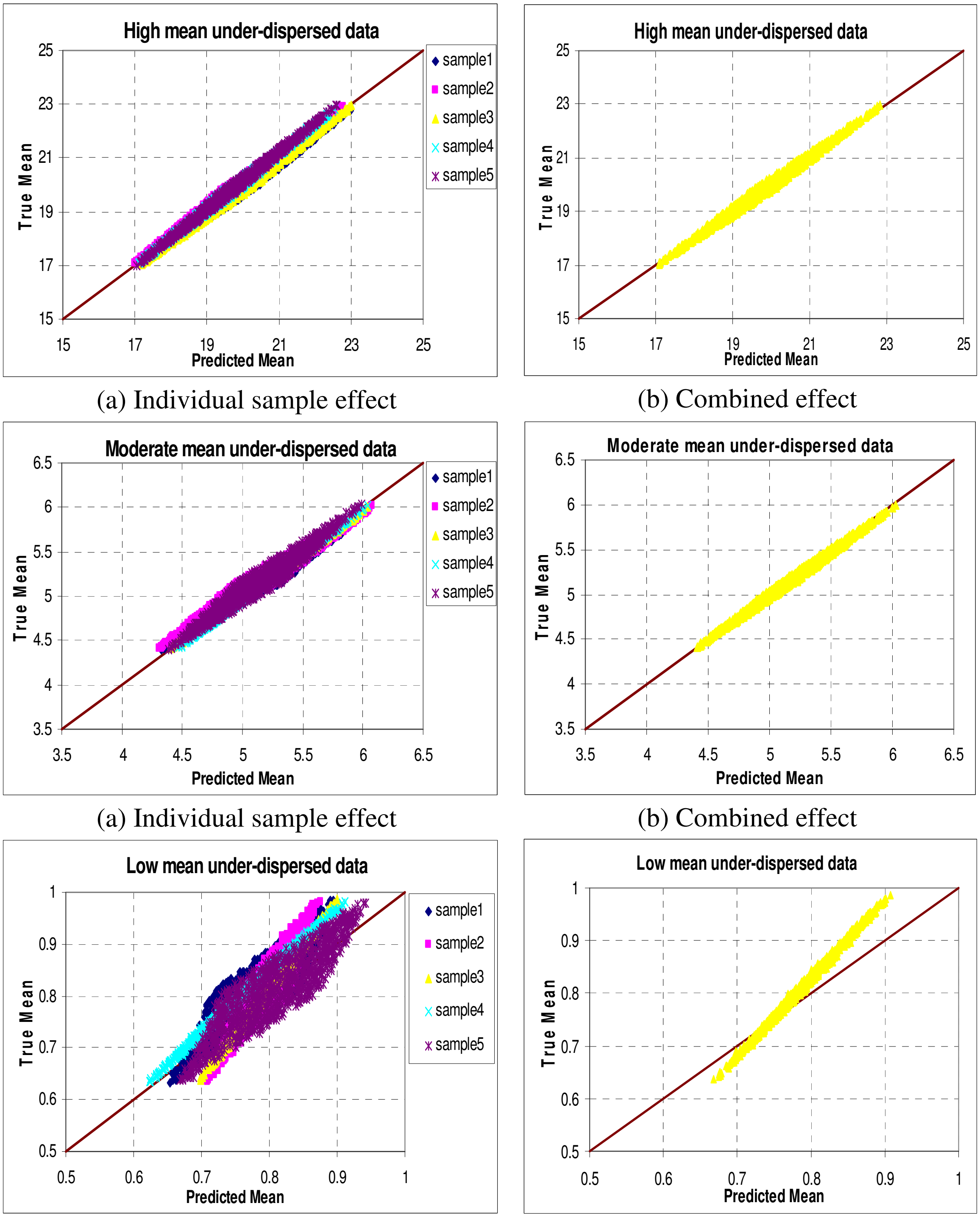
(a) Individual sample effect

(b) Combined effect

(a) Individual sample effect

(b) Combined effect

Figure 3.7: Prediction accuracy for over-dispersed datasets

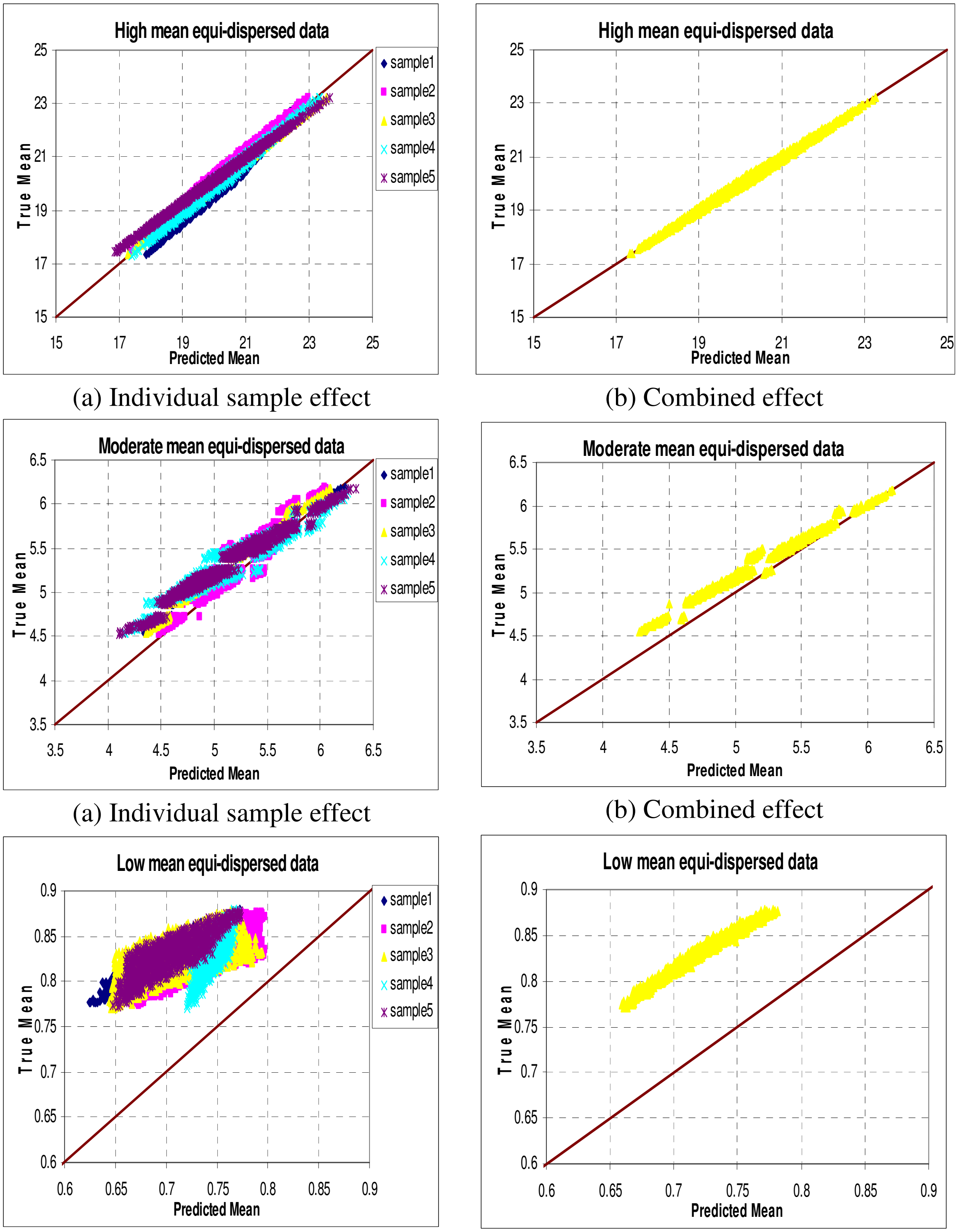


(a) Individual sample effect (b) Combined effect

(a) Individual sample effect (b) Combined effect

(a) Individual sample effect (b) Combined effect

Figure 3.8: Prediction accuracy for under-dispersed datasets



(a) Individual sample effect

(b) Combined effect

(a) Individual sample effect

(b) Combined effect

(a) Individual sample effect

(b) Combined effect

Figure 3.9: Prediction accuracy for equi-dispersed datasets

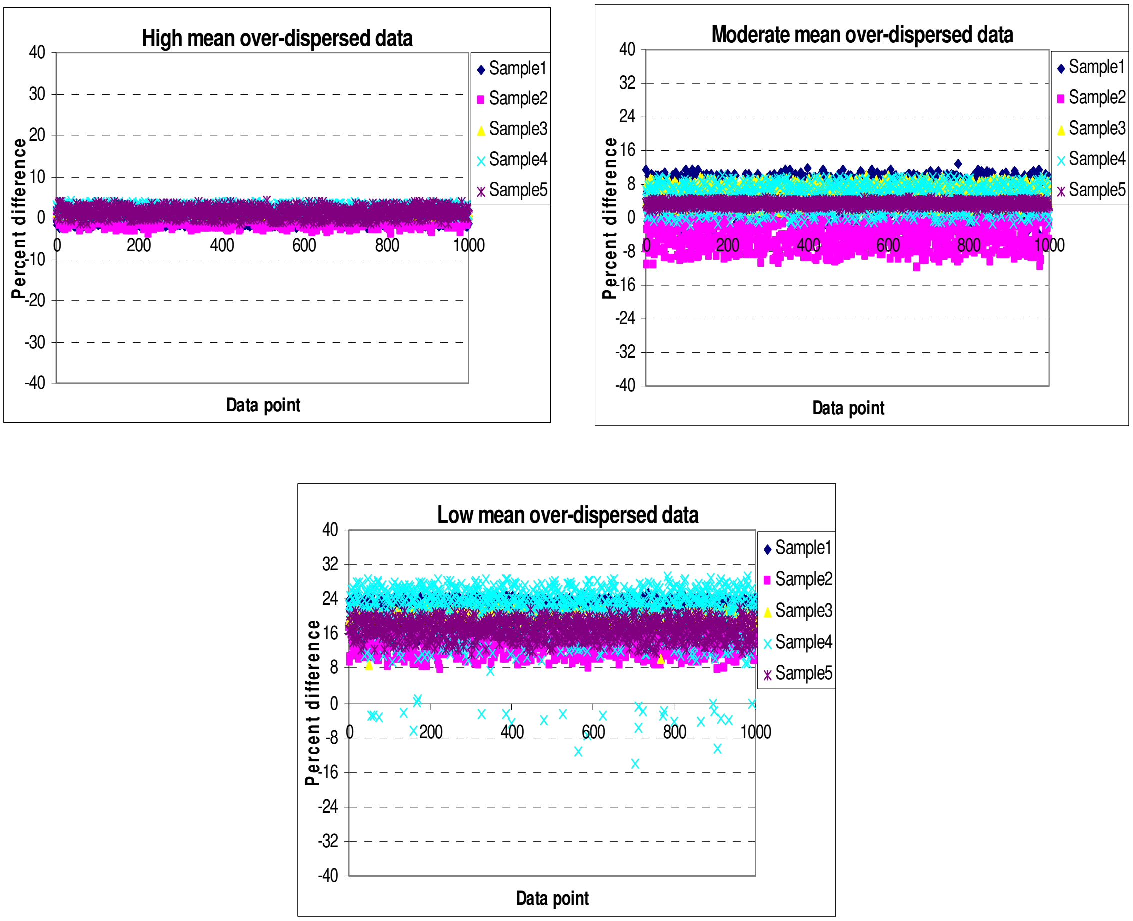


Figure 3.10: Percent difference between true mean and predicted mean for over-dispersed datasets

Figure 3.11 gives the percent difference of the equi-dispersed datasets. As seen, the performance of COM-Poisson distribution becomes worse from high mean to low mean for all equi-dispersed datasets.

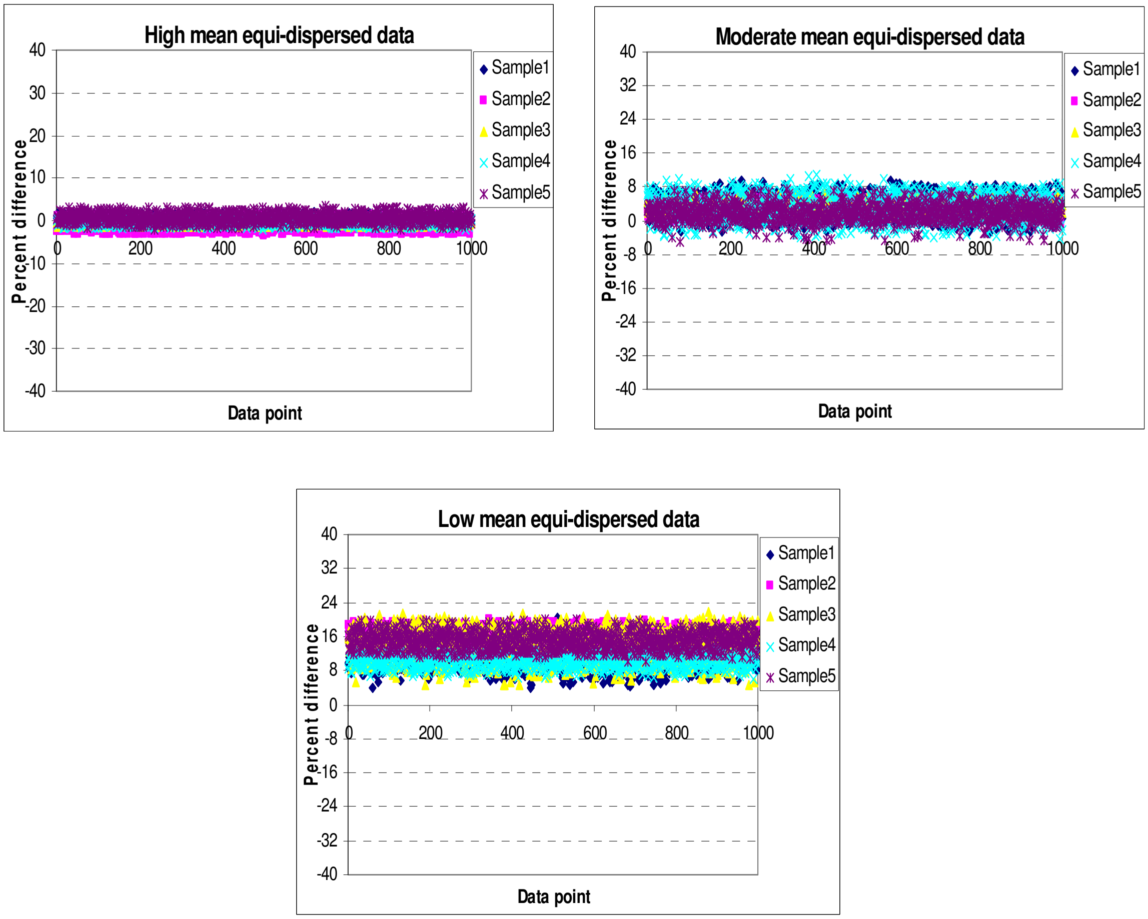


Figure 3.11: Percent difference between true mean and predicted mean for equi-dispersed datasets

The percent difference for the under-dispersed datasets is given in Figure 3-12. The COM-Poisson performs well for under-dispersed data when compared to over-dispersed and equi-dispersed datasets. Similar to other dispersions, the COM-Poisson performs well for high mean datasets and worse for low mean datasets.

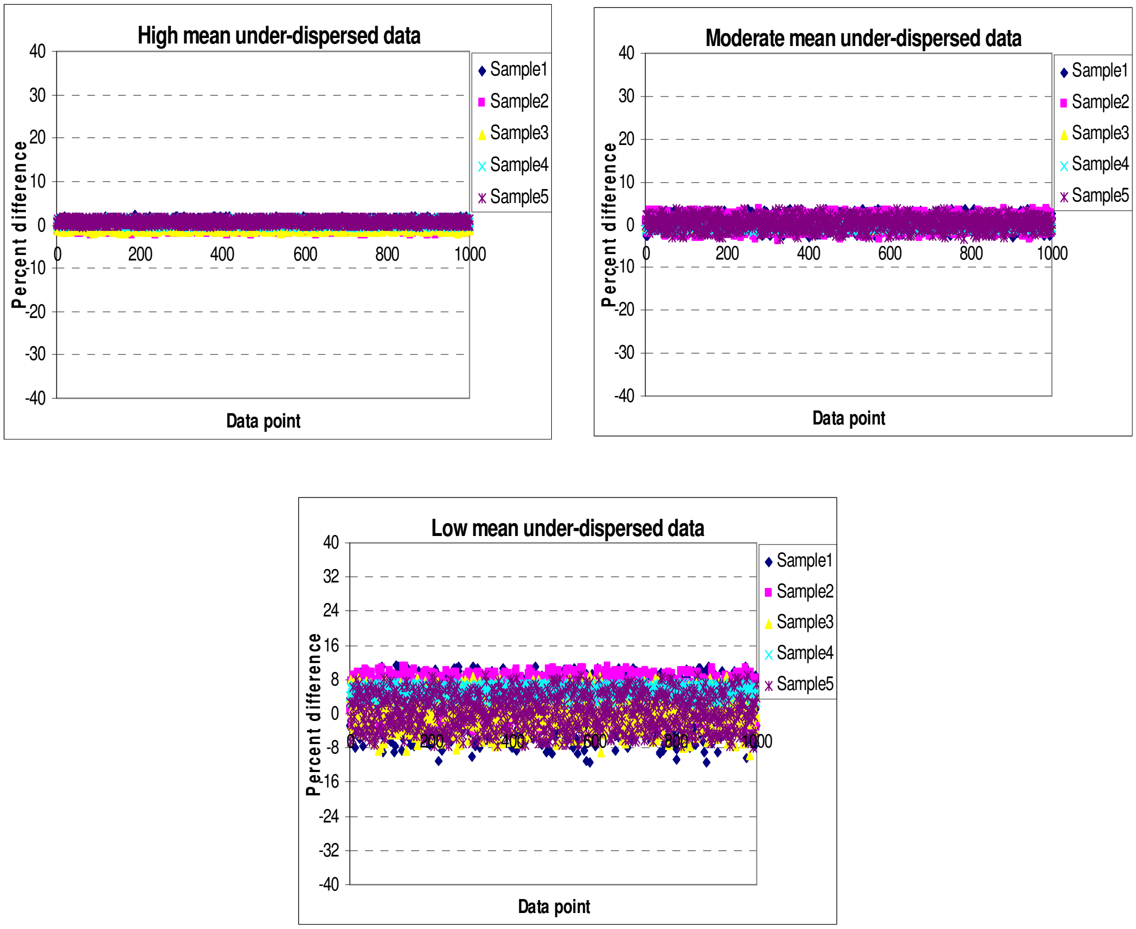


Figure 3.12: Percent difference between true mean and predicted mean for under-dispersed datasets

3.2.5 Accuracy of the asymptotic mean approximation

The centering parameter μ is believed to adequately approximate the mean when μ >10 based on the asymptotic approximation developed by Shmueli et al. (2005). However, the deviation of μ for mean values below 10 (μ < 10) has not been investigated. One sample from each of the nine scenarios was chosen as a basis for estimating the accuracy of the asymptotic mean approximation. First, the μ and v parameters were calculated from the estimated parameters. The goodness of this approximation was examined by simulating 100,000 random values from the COM-Poisson for a given μ and v. The mean of the simulated values against the asymptotic mean approximation ($E[Y] \approx \mu + 1/2\nu - 1/2$) was then plotted.

The results showed that the asymptotic mean approximates the true mean accurately even for $10 > E[Y] > 5$. As the sample mean value decreases below 5, the accuracy of the approximation drops. As seen in Figure 3.13, the asymptotic approximation holds well for all datasets with high and moderate mean values irrespective of the dispersion in the data. The approximation is also accurate for low sample mean values for under-dispersed datasets. The accuracy drops significantly for the low sample mean values for over-dispersed and equi-dispersed datasets. There is not much difference between the asymptotic mean approximation and the true mean for $E[Y] > 10$. It starts to deviate at the moderate mean values although the difference is not high. The difference can clearly be observed for the low sample mean values, particularly for the over-dispersed and equi-dispersed datasets. This shows that one must be careful in using the asymptotic approximation for the mean of the COM GLM to estimate future event counts for datasets characterized by low sample mean values.

3.2.6 Computational time

The computational time needed for the WinBUGS MCMC implementation of the COM GLM was also investigated. The COM-Poisson MCMC model was ran on a computer with a 1.5GHz Pentium 4 processor and 512 MB of RAM. Each run consisted of 3 chains of 50,000 replications each. The computational times for all of the datasets were plotted against the mean values of the counts in those datasets and are shown in the Figure 3.14. Datasets with higher sample means required more computational time for a given number of replications than datasets with low sample mean. This is mainly attributable to the convergence of the '*S*' term. The centering parameter causes the numerator to be large for high sample mean values, requiring that more terms be included in the approximation to achieve suitable convergence of the approximation of the series. Also, it is important to note that the over-dispersed datasets required more computational time than the other type of datasets. The shape parameter plays vital role in the convergence of the *S* term. Lower values for the shape parameter require higher amounts of computational time. All the computations are performed using a computer with a 1.50 GHz Pentium 4 CPU and

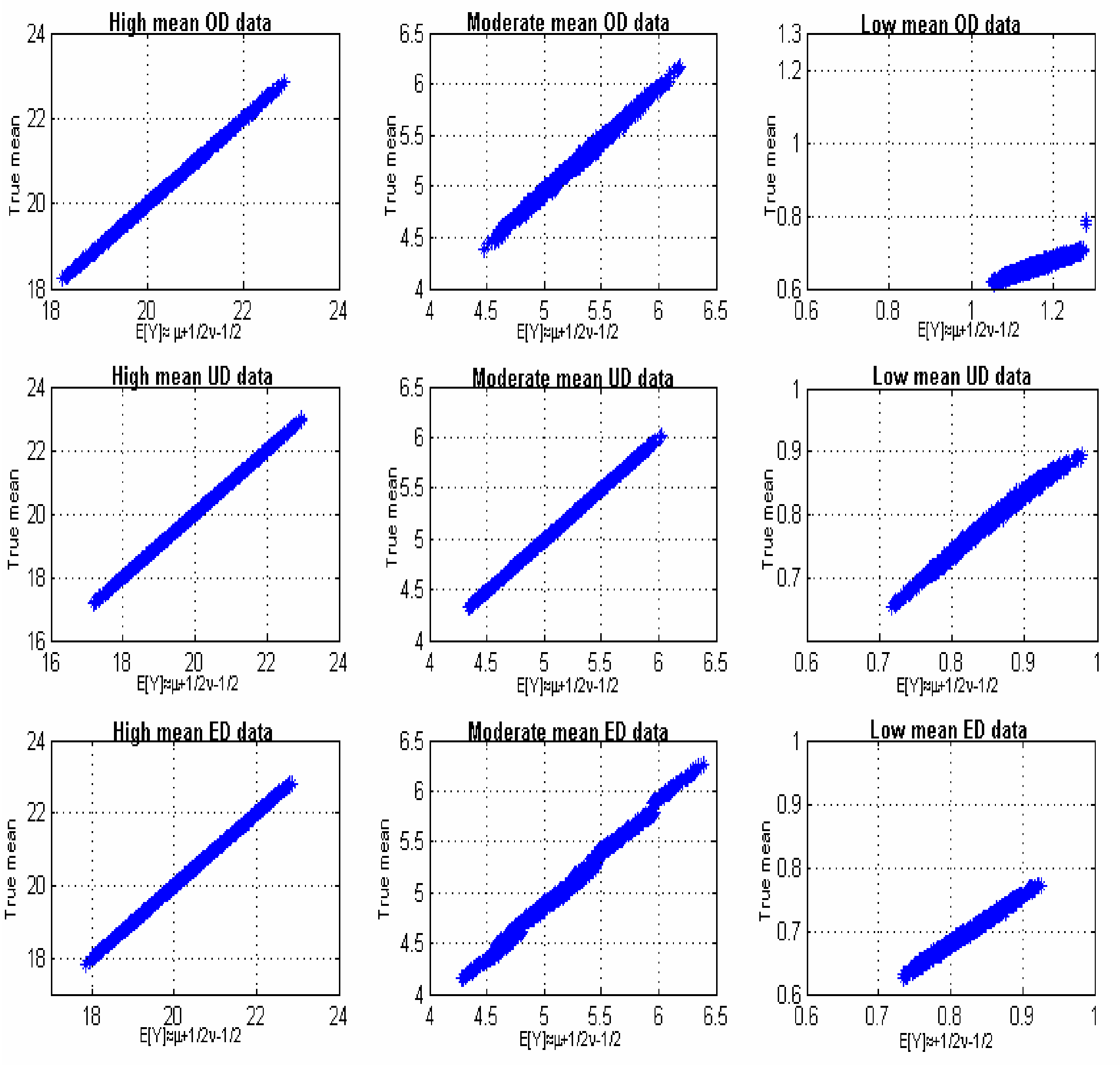


Figure 3.13: Mean versus asymptotic approximation

512 MB of RAM. However, using different computers with more processing speed and RAM would cut the time by two thirds.

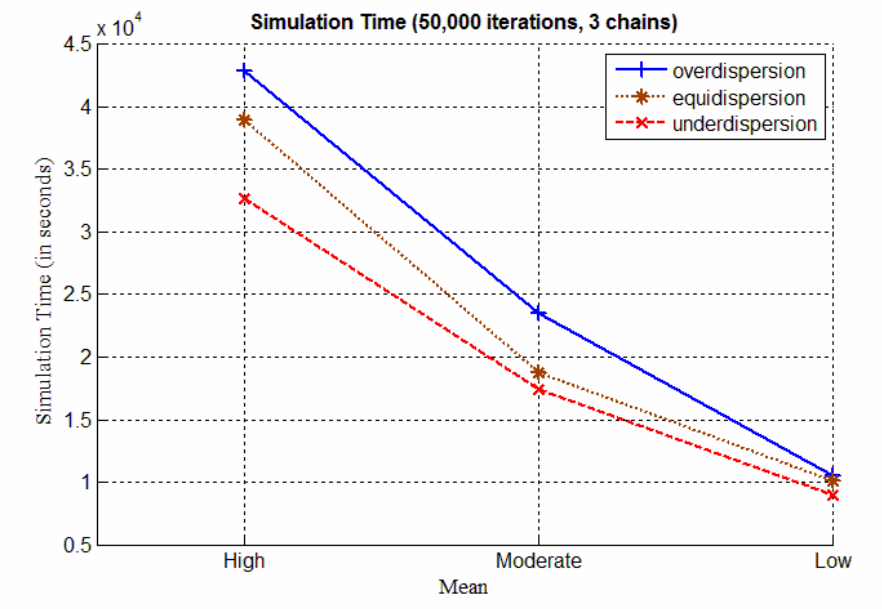


Figure 3.14: Computational time for the WinBUGS MCMC implementation

3.3 Discussion

This study shows that the COM-Poisson GLM is flexible in handling count data irrespective of the dispersion in the data. The following results are drawn from this analysis:

- First, the true parameters lie in the 95% credible interval for nearly all cases and are generally close to the estimated posterior mean of the parameters. The credible intervals were found to be wider for the low mean values for both the centering and shape parameters. The bias in the prediction of the parameters and the mean also increases as the data sample mean values decreases. Even at the low sample mean values, the bias is considerably less for under-dispersed datasets than for over-dispersed and equi-dispersed datasets. Despite its flexibility in handing count data with all dispersions, the COM-Poisson distribution suffers from important limitations for low mean over-dispersed data. This similar behavior is also exhibited by Negative Binomial (Poisson-gamma) models (Lord, 2006).

- Second, the asymptotic approximation of the mean suggested by Shmueli et al (2005) approximates the true mean adequately for $E[Y] > 5$. This value found through numerical analysis of the COM-Poisson GLM is substantially lower than the lower bound value of 10 suggested by Shmueli et al. (2005). As the sample mean value decreases, the accuracy of the approximation becomes lower. The asymptotic approximation is accurate for all datasets with high and moderate sample mean values irrespective of the dispersion in the data. The approximation is also accurate for low sample mean values for under-dispersed datasets. However, the accuracy drops substantially for low sample mean values for over-dispersed and equi-dispersed datasets.

- Third, datasets with higher sample mean values required more computational time for a given number of replications than the low mean datasets did. Similarly, it is important to note that the over-dispersed datasets required more computational time than the other type of datasets.

3.3 Summary

There exist various distributions for analyzing the count data. Out of all, only few distributions have the capability of handing under-dispersed and over-dispersed datasets and the COM-Poisson distribution is one among them. This chapter has documented the performance of COM-Poisson GLM for datasets characterized by different variances and sample mean values.

The first section gave a brief methodology about data simulation and testing protocol. The second section presented the results concerning the error term, prediction accuracy, accuracy of the asymptotic approximation of the mean and computational effort required for the MCMC implementation of the COM-Poisson GLM. The results of this study showed that the COM-Poisson GLMs can handle under-, equi- and over-dispersed datasets with different mean values, although the credible intervals are found to be wider for low sample mean values. Despite its limitations for low sample mean values for over-dispersed datasets, the COM-Poisson GLM is still a flexible method for analyzing count data. The asymptotic approximation of the mean is accurate for all datasets with high and moderate sample mean values irrespective of the dispersion in the data, and it is also accurate for low sample mean values for under-dispersed datasets. Furthermore, the computational effort required for the MCMC implementation of the COM-Poisson GLM is not prohibitive. The last section presented a brief discussion about the results. Finally, from the results of this chapter, it is concluded that the COM-Poisson GLM is a promising, flexible regression model for count data. The next chapter describes the application of the COM-Poisson GLM for analyzing crash data exhibiting over-dispersion.

CHAPTER IV

APPLICATION OF THE COM-POISSON GLM TO TRAFFIC CRASH DATA EXHIBITING OVER-DISPERSION [*]

It is clear from the results of Chapter III that the COM-Poisson GLM is a promising, flexible regression model for count data. Also from the literature review in Chapter II, it is seen that several studies have documented important limitations associated with Poisson and NB models. Thus, there is a need to evaluate whether COM-Poisson models could be used for modeling motor vehicle crashes.

The objectives of this chapter are to evaluate the application of the COM-Poisson GLM for analyzing motor vehicle crashes and compare the results with those produced from the NB model. Nobody has so far examined how the COM-Poisson GLM could be used for modeling crash data using common functional forms linking crash data to traffic flow variables (often referred to as general Annual Average Daily Traffic or AADT models). Although traffic-flow only models could suffer from omitted variables bias, they are still the most popular type of models developed and used by transportation safety analysts (Hauer, 1997; Persaud et al., 2001). They are often preferred over models that include several covariates because they can be easily re-calibrated when they are developed in one jurisdiction and applied to another (Persaud et al., 2001; Lord and Bonneson, 2005). In fact, this type of model will be the kind of model used for estimating the safety performance of rural and urban highways as well as for intersections in the forthcoming Highway Safety Manual (HSM) (Hughes et al., 2005).

[*] Part of this chapter is reprinted with permission from “Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes” by Lord, D., Guikema, S., Geedipally, S.R., 2008. Accident Analysis & Prevention, 40 (3), 1123-1134. Copyright [2008] by Elsevier Ltd.

This chapter is organized as follows: The first section presents the summary statistics of the two datasets. The second section summarizes the results of the comparison analysis. The third section provides a discussion about the application of COM-Poisson GLMs in highway safety. The last section presents a summary of the analysis carried out in this chapter.

4.1 Methodology

This section describes the methodology used for estimating and comparing the two types of model. For each dataset, COM-Poisson GLMs and NB models were initially estimated using the entire dataset. Then, five samples, which consisted of 80% of the original data, were randomly extracted. The models were developed using the subsets and were then applied to the remaining 20% to evaluate their predictive performance.

The functional form used for models were the following (see Eq. (2.2 & 2.7) above):

$$\text{Toronto intersection data: } \mu_i = \beta_0 F_{Maj\_i}^{\beta_1} F_{Min\_i}^{\beta_2} \tag{4.1}$$

$$\text{Texas segment data: } \mu_j = \beta_0 L F_j^{\beta_1} \tag{4.2}$$

where,

μ_i = the mean number of crashes per year for intersection i;

μ_j = the mean number of crashes per year for segment j;

F_{Maj_i} = entering flow for the major approach (average annual daily traffic or AADT) for intersection i;

F_{Min_i} = entering flow for the minor approach (average annual daily traffic or AADT) for intersection i;

F_j = flow traveling on segment j (average annual daily traffic or AADT) and time period t;

L = length in miles for segment j; and,

$\beta_0, \beta_1, \beta_2$ = estimated coefficients.

The functional forms described above are very frequently used by transportation safety analysts. Although they are not considered the most adequate functional form (see Miaou and Lord, 2003), since they under-perform near the boundary conditions (at least for intersections), they are still relevant for this study, as they are considered established functional forms in the highway safety literature.

Several methods were used for estimating the GOF and predictive performance of the models. The methods used in this research include the following:

4.1.1 Deviance information criterion (DIC)

The DIC is defined as

$$DIC = \bar{D} + p_D \tag{4.3}$$

where $\bar{D} = -2\log L$ represents the posterior mean of the deviance of the un-standardized model where L is the mean of the model log likelihood; $p_D = \bar{D} - D(y|\bar{\theta})$ represents the penalty for the number of effective model parameters where $D(y|\bar{\theta})$ is the point estimate of deviance for the posterior means $\bar{\theta}$.

4.1.2 Mean absolute deviance (MAD)

MAD provides a measure of the average mis-prediction of the model (Oh et al, 2003). It is computed using the following equation:

$$\text{Mean Absolute Deviance (MAD)} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{4.4}$$

4.1.3 Mean squared predictive error (MSPE)

MSPE is typically used to assess the error associated with a validation or external data set (Oh et al., 2003). It can be computed using Equation (4.5):

$$\text{Mean Squared Predictive Error (MSPE)} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \qquad (4.5)$$

The coefficients of the COM-Poisson GLMs and FB NB models were estimated using the software WinBUGS (Spiegelhalter et al., 2003) and the coefficients of MLE NB models were estimated using the software SAS (SAS, 2002) (note: to distinguish between the Bayesian and the MLE methods, FB NB refers to a model estimated using a Bayesian approach while MLE NB refers to a model estimated using the Frequentist approach). Vague or non-informative hyper-priors were utilized for the COM-Poisson GLMs and FB NB (described below). A total of 3 Markov chains were used in the model estimation process with 35,000 iterations per chain and the thinning parameter was set to 1. The first 20,000 iterations (burn-in samples) were discarded. Thus, the remaining 15,000 iterations were used for estimating the coefficients. The Gelman-Rubin (G-R) convergence statistic was used to verify that the simulation runs converged properly. In the analysis, the G-R statistic was less than 1.1. For comparison, Mitra and Washington (2007) suggested that convergence was achieved when the G-R statistic was less than 1.2.

4.2 Data Description

To accomplish the objectives of this chapter, NB and COM-Poisson GLMs were developed and compared using two datasets. The first dataset is an urban 4-legged signalized intersection data set collected from Toronto, Ontario for the year 2005. This dataset had been used in many studies so far (Lord, 2000; Miaou and Lord, 2003). As stated in Miaou and Lord (2003), crashes in this dataset include both intersection and intersection-related crashes as reported by the police that are located within about 15 m (50 ft) from the center of the intersection and does not include crashes involving pedestrians, animals, and cyclists. This dataset consisted of 868 4-legged signalized intersections with a total of 10030 crashes for the year 1995. To evaluate the performance of COM for different datasets with varying properties, 5 random samples each of 694 intersections (80% of total data) were collected for fitting the COM-Poisson and NB models. The remaining samples of 174 intersections were used as the predicting data so as to know the performance of the models for predicting crashes for a new dataset. The properties of the Toronto data set and the random samples are given in Table 4.1.

The second dataset constituted of crashes on Texas 4-lane undivided and divided segments from the year 1997 to 2001. Only the crashes that are non-intersection related were considered in this analysis. There were 3220 segments which are used in this study. Similar to the first dataset, 5 different random samples each of 2576 segments were collected for fitting the COM-Poisson and NB models. The remaining 644 segments were used to know the performance of the models for predicting crashes for a new dataset. The detailed description of each of the dataset is given in Table 4.2. Although there are many variables that influence the occurrence of crashes, only traffic flow is considered in this analysis.

4.3 Modeling Results

This section presents the modeling results for the COM-Poisson GLMs as well as for the FB and MLE NB models and is divided into three parts. The first part explains the modeling results for the Toronto data. The second part provides details about the modeling results for the Texas data. The last part documents the marginal analysis used for examining the regression coefficients.

4.3.1 Toronto data

Table 4.3 summarizes the results of the COM-Poisson GLMs for the Toronto data. This table shows that the coefficients for the flow parameters are below one, which indicates that the crash risk increases at a decreasing rate as traffic flow increases. It should be pointed out that the 95% marginal posterior credible intervals for each of the coefficients did not include the origin.

Table 4.1: Description of variables for Toronto dataset and random samples

		Fitting data				Predicting data			
		Min.	Max.	Average	Total	Min.	Max.	Average	Total
Full data	Crashes	0	54	11.56 (10.02)	10030	--	--	--	--
	Major AADT	5469	72178	28044.81 (10660.4)	--	--	--	--	--
	Minor AADT	53	42644	11010.18 (8599.40)	--	--	--	--	--
Sample1	Crashes	0	54	11.49 (9.94)	7974	0	51	11.82 (10.33)	2056
	Major AADT	5469	72178	28097.36 (10656.9)	--	9622	67214	27835.21 (10702.5)	--
	Minor AADT	71	41288	10904.03 (8421.3)	--	53	42644	11433.55 (9289.40)	--
Sample2	Crashes	0	54	11.50 (9.96)	7983	0	48	11.76 (10.28)	2047
	Major AADT	5469	67214	27946.05 (10490.5)	--	7361	72178	28438.69 (11335.8)	--
	Minor AADT	53	42644	10862.04 (8532.00)	--	877	41029	11601.03 (8863.55)	--
Sample3	Crashes	0	53	11.67 (10.07)	8099	0	54	11.10 (9.84)	1931
	Major AADT	5469	72178	28115.98 (10824.1)	--	5967	56623	27760.95 (10005.6)	--
	Minor AADT	71	42644	11169.03 (8678.03)	--	53	36002	10376.61 (8272.24)	--
Sample4	Crashes	0	54	11.63 (10.02)	8074	0	50	11.24 (10.02)	1956
	Major AADT	5469	68594	28072.28 (10652.6)	--	7361	72178	27935.25 (10721.7)	--
	Minor AADT	465	42644	11119.84 (8749.1)	--	53	41288	10572.82 (7983.41)	--
Sample5	Crashes	0	54	11.69 (10.08)	8113	0	44	11.02 (9.79)	1917
	Major AADT	5469	72178	28103.95 (10641.9)	--	5967	56697	27808.91 (10761.52)	--
	Minor AADT	71	42644	11104.99 (8712.5)	--	53	34934	10632.03 (8145.70)	--

Table 4.2: Description of variables for Texas dataset and random samples (1997-2001)

		Fitting data				Predicting data			
		Min.	Max.	Average	Total	Min.	Max.	Average	Total
Full data	Crashes	0	108	4.89 (8.45)	15753	--	--	--	--
	AADT	42	89264	8639.27 (6606.57)	--	--	--	--	--
	Length (miles)	0.1	11.21	0.80 (1.02)	2576.18	--	--	--	--
Sample1	Crashes	0	108	4.94 (8.75)	12722	0	52	4.71 (7.16)	3031
	AADT	42	89264	8704.98 (6715.86)	--	420	52294	8376.39 (6147.98)	--
	Length (miles)	0.1	11.21	0.80 (1.02)	2054.91	0.1	8.319	0.81 (1.04)	521.269
Sample2	Crashes	0	108	4.93 (8.72)	12703	0	48	4.74 (7.31)	3050
	AADT	42	89264	8699.58 (6681.18)	--	266	52294	8398.03 (6298.54)	--
	Length (miles)	0.1	11.21	0.80 (1.03)	2063.87	0.1	8.517	0.80 (0.96)	512.313
Sample3	Crashes	0	108	4.91 (8.21)	12655	0	97	4.81 (9.38)	3098
	AADT	158	89264	8645.19 (6685.83)	--	42	53714	8615.58 (6284.50)	--
	Length (miles)	0.1	8.548	0.81 (1.02)	2076.86	0.1	11.21	0.78 (1.01)	499.318
Sample4	Crashes	0	108	4.94 (8.44)	12731	0	97	4.69 (8.53)	3022
	AADT	42	89264	8646.10 (6619.57)	--	158	53714	8611.93 (6559.38)	--
	Length (miles)	0.1	11.21	0.80 (1.01)	2073.67	0.1	8.517	0.78 (1.04)	502.51
Sample5	Crashes	0	108	5.00 (8.69)	12891	0	53	4.44 (7.43)	2862
	AADT	42	89264	8700.86 (6712.74)	--	264	52294	8392.89 (6162.47)	--
	Length (miles)	0.1	8.548	0.81 (1.02)	2091.04	0.1	11.21	0.75 (1.02)	485.143

Table 4.3: Modeling results for the COM-Poisson GLMs using the Toronto data

Estimates[†]	Full data	Sample1	Sample2	Sample3	Sample4	Sample5	Average
$Ln(\beta_0)$	-11.53 (0.4159)[‡]	-11.7589 (0.742)	-11.7252 (0.560)	-11.2626 (0.487)	-11.2033 (0.729)	-11.1643 (0.709)	--
β_1	0.6350 (0.04742)	0.6527 (0.076)	0.6641 (0.055)	0.6078 (0.049)	0.6071 (0.072)	0.5949 (0.066)	--
β_2	0.795 (0.03101)	0.7999 (0.029)	0.7854 (0.029)	0.7960 (0.031)	0.7912 (0.030)	0.8010 (0.027)	--
ν	0.3408 (0.02083)	0.3333 (0.023)	0.3454 (0.023)	0.3359 (0.023)	0.3497 (0.024)	0.3499 (0.024)	--
DIC	4953.7	3974.34	3953.69	3981.33	3953.66	3956.85	--
MAD_{fit}	4.129	4.141	4.075	4.156	4.132	4.074	4.118
$MSPE_{fit}$	33.664	34.433	33.102	34.108	33.508	33.176	33.665
MAD_{pred}	--	4.082	4.3003	4.034	4.106	4.316	4.168
$MSPE_{pred}$	--	30.529	34.695	32.339	34.059	34.663	33.257

[†] The coefficient estimates are based on the mode (posterior value) (see discussion above)

[‡] Posterior credible standard error

Table 4.4 summarizes the results of the FB NB models for the Toronto data. This table exhibits similar characteristics as for the COM-Poisson GLMs in terms of GOF statistics and predictive performance despite the fact that the coefficients are a little bit different. Nonetheless, this difference did not affect the fit and predictive capabilities of the models. A comparison of the models' output is presented below.

Table 4.5 summarizes the results of the MLE NB models for the Toronto data. This table shows exactly the same results as for the FB NB. This is expected since a vague prior was used for the FB NB models. The results indicate that the FB NB models are relatively stable and can, therefore, be compared with the COM-Poisson GLM.

Table 4.4: Modeling results for the FB NB models using the Toronto data

Estimates	Full data	Sample1	Sample2	Sample3	Sample4	Sample5	Average
Ln(β_0)	-10.11 (0.4794)	-9.862 (0.4018)	-10.48 (0.4359)	-9.799 (0.5509)	-10.14 (0.5256)	-9.87 (0.5272)	--
β_1	0.6071 (0.046)	0.5788 (0.039)	0.6459 (0.044)	0.5775 (0.052)	0.6057 (0.051)	0.5787 (0.055)	--
β_2	0.6852 (0.021)	0.6903 (0.022)	0.684 (0.025)	0.6848 (0.023)	0.6902 (0.024)	0.6918 (0.023)	--
ϕ	7.12 (0.619)	6.898 (0.669)	7.256 (0.721)	7.045 (0.687)	7.388 (0.734)	7.567 (0.756)	--
DIC	4777.59	3821.9	3817.8	3836.35	3811.16	3824.74	--
MAD_{fit}	4.141	4.174	4.094	4.168	4.145	4.096	4.136
MSPE_{fit}	32.742	33.503	32.104	33.271	32.527	32.354	32.750
MAD_{pred}	--	4.024	4.379	4.058	4.121	4.346	4.186
$\text{MSPE}_{\text{pred}}$	--	29.594	35.091	30.855	33.331	33.989	32.572

Figure 4.1 compares the estimated number of crashes from the COM-Poisson and NB models for three minor AADT flows (F_{Min}). The figure illustrates that the estimated values are slightly different, especially when F_{Min} is equal to 500 veh/day, with the COM-Poisson output being always lower than the NB output. For $F_{\text{Min}} = 500$, the maximum absolute difference is about 0.9 crash per year. At the other end of the spectrum, the maximum absolute difference is about 2 crashes per year for $F_{\text{Maj}} = 70000$ and $F_{\text{Min}} = 3000$. Although this difference appears to be large, the relative difference is about 17%. It should be pointed out that when the posterior mean value is used for the COM-Poisson model rather than the centering parameter μ (e.g., assuming μ is the predicted mean), both curves get closer for all minor flow values. For $F_{\text{Min}} = 3000$, the absolute maximum difference becomes less than 1 crash per year. Thus, both models could be used for analyzing this dataset.

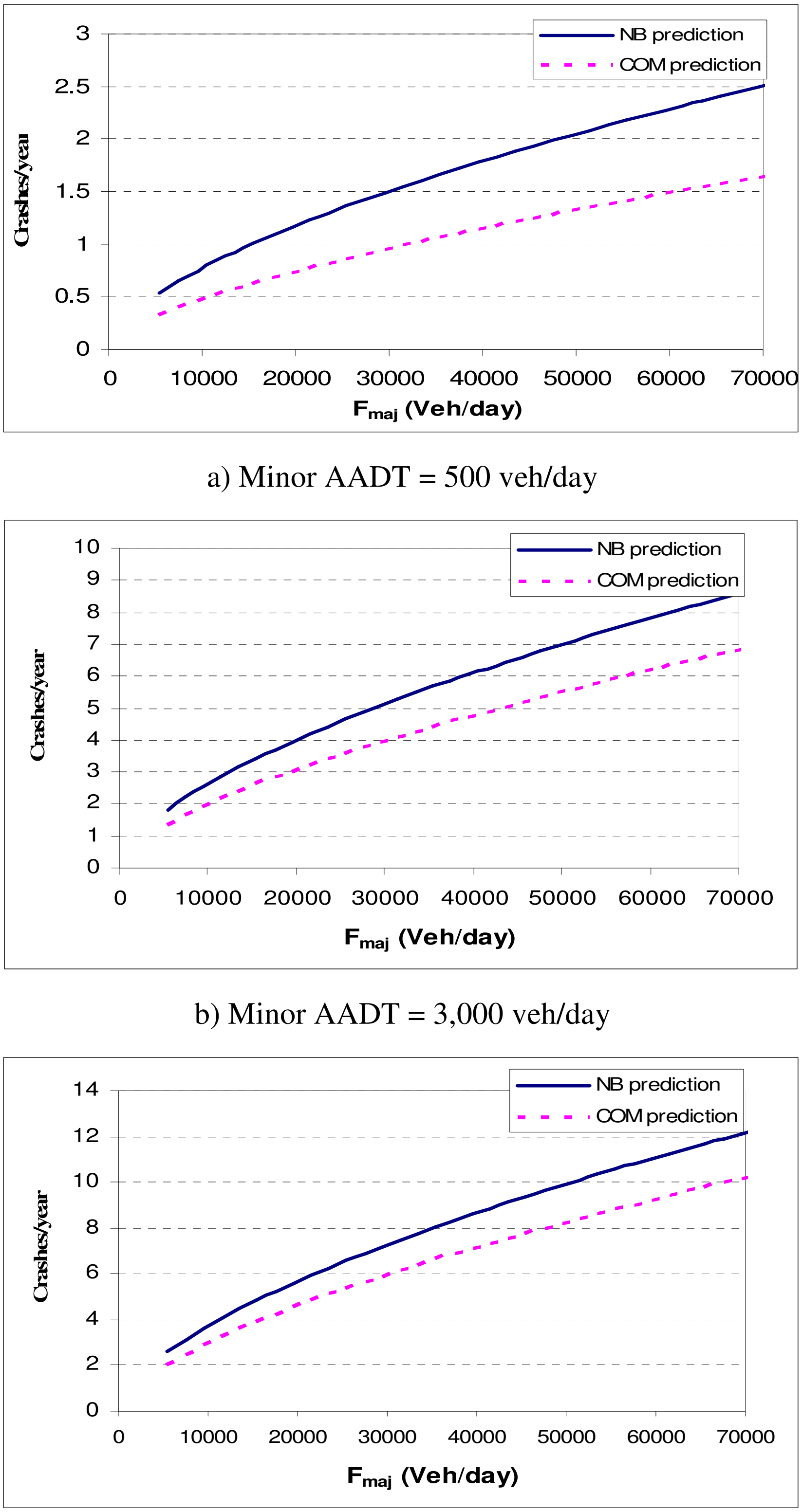
Table 4.5: Modeling results for the MLE NB models using the Toronto data

Estimates	Full data	Sample1	Sample2	Sample3	Sample4	Sample5	Average
Ln(β_0)	-10.2458 (0.465)	-10.1664 (0.525)	-10.4031 (0.520)	-9.7398 (0.513)	-10.2040 (0.513)	-9.8473 (0.511)	--
β_1	0.6207 (0.046)	0.6079 (0.051)	0.6393 (0.051)	0.5707 (0.051)	0.6119 (0.051)	0.5778 (0.051)	--
β_2	0.6853 (0.0211)	0.6910 (0.0241)	0.6826 (0.024)	0.6860 (0.024)	0.6905 (0.024)	0.6903 (0.0234)	--
$\alpha^{\dagger}$	0.1398 (0.0122)	0.1443 (0.014)	0.1372 (0.014)	0.1410 (0.014)	0.1349 (0.0134)	0.1315 (0.0134)	--
AIC	5077.3	4068.8	4052.6	4080.5	4045.2	4054.3	--
MAD_{fit}	4.142	4.170	4.092	4.168	4.146	4.096	4.136
$MSPE_{fit}$	32.699	33.444	32.127	33.264	32.517	32.370	32.737
MAD_{pred}	--	4.026	4.374	4.062	4.117	4.348	4.185
$MSPE_{pred}$	--	29.547	34.973	30.898	33.271	34.002	32.538

[†] Note: $\alpha = {}^{1}\!/\!{}_{\phi}$

4.3.2 Texas data

Table 4.6 summarizes the results of the COM-Poisson models for the Texas data. Similar to the first dataset, the 95% marginal posterior credible intervals for each of the coefficients did not include the origin. In addition, the coefficients do not vary significantly between the different samples.



a) Minor AADT = 500 veh/day

b) Minor AADT = 3,000 veh/day

c) Minor AADT = 5,000 veh/day

Figure 4.1: Estimated values (crashes/year) for the Toronto data: NB and COM-Poisson models

Table 4.6: Modeling results for the COM-Poisson GLMs using the Texas data

Estimates[†]	Full data	Sample1	Sample2	Sample3	Sample4	Sample5	Average
Ln(β_0)	-8.235 (0.242)[‡]	-8.442 (0.267)	-8.333 (0.284)	-7.877 (0.223)	-8.155 (0.234)	-8.338 (0.249)	--
β_1	1.081 (0.025)	1.102 (0.028)	1.089 (0.030)	1.044 (0.023)	1.074 (0.025)	1.092 (0.026)	--
ν	0.3608 (0.012)	0.3504 (0.014)	0.3465 (0.013)	0.3699 (0.014)	0.3701 (0.015)	0.36 (0.013)	--
DIC	13325.6	10688.8	10711.2	10673.9	10652.6	10710.6	--
MAD_{fit}	2.385	2.433	2.435	2.369	2.371	2.415	2.401
$MSPE_{fit}$	21.985	24.297	23.708	18.970	20.050	22.938	21.991
MAD_{pred}	--	2.240	2.242	2.388	2.413	2.283	2.313
$MSPE_{pred}$	--	14.462	16.650	31.748	28.745	18.835	22.088

[†] The coefficient estimates are based on the mode (posterior value) (see discussion above)

[‡] Posterior credible standard error

Table 4.7 summarizes the results of the FB NB models for the Texas data. This table indicates that FB NB models estimate a slightly lower value for the coefficient for the traffic flow variable than for the COM-Poisson GLMs. Similar to the Toronto data, the COM-Poisson GLMs offer the same statistical performance as for FB NB models.

Table 4.8 summarizes the results of the MLE NB models for the Texas data. This table shows exactly the same results as for the FB NB.

Figure 4.2 shows the comparison results between the estimated number of crashes per mile per 5-year of the COM-Poisson and NB models for the Texas data (full dataset). This figure illustrates that both estimates are indeed very close.

Table 4.7: Modeling results for the FB NB models using the Texas data

Estimates	Full data	Sample1	Sample2	Sample3	Sample4	Sample5	Average
$Ln(\beta_0)$	-6.512 (0.227)	-6.701 (0.264)	-6.488 (0.211)	-6.356 (0.221)	-6.449 (0.209)	-6.618 (0.204)	--
β_1	0.9206 (0.025)	0.9403 (0.029)	0.9178 (0.023)	0.9046 (0.024)	0.9134 (0.023)	0.932 (0.022)	--
ϕ	3.229 (0.177)	3.155 (0.189)	3.073 (0.184)	3.235 (0.198)	3.253 (0.198)	3.328 (0.200)	--
DIC	12408.7	9882.13	9908.99	9988.68	9964.02	9983.94	--
MAD_{fit}	2.437	2.466	2.508	2.430	2.424	2.453	2.453
$MSPE_{fit}$	20.699	22.758	22.487	18.070	18.284	21.651	20.658
MAD_{pred}	--	2.297	2.143	2.509	2.464	2.378	2.358
$MSPE_{pred}$	--	12.929	13.307	31.162	29.854	17.369	20.924

Table 4.8: Modeling results for the MLE NB models using the Texas data

Estimates	Full data	Sample1	Sample2	Sample3	Sample4	Sample5	Average
$Ln(\beta_0)$	-6.5605 (0.199)	-6.6293 (0.224)	-6.4570 (0.224)	-6.4266 (0.222)	-6.4652 (0.220)	-6.6163 (0.220)	--
β_1	0.9260 (0.022)	0.9324 (0.025)	0.9143 (0.025)	0.9125 (0.025)	0.9151 (0.024)	0.9318 (0.024)	--
$\alpha^{\dagger}$	0.3095 (0.017)	0.3172 (0.019)	0.3255 (0.020)	0.3094 (0.019)	0.3075 (0.019)	0.3009 (0.018)	--
AIC	13375	10674	10724	10773	10741	10749	--
MAD_{fit}	2.440	2.463	2.506	2.434	2.424	2.453	2.453
$MSPE_{fit}$	20.766	22.654	22.439	18.190	18.289	21.647	20.664
MAD_{pred}	--	2.293	2.141	2.510	2.463	2.378	2.357
$MSPE_{pred}$	--	12.856	13.282	31.146	29.861	17.365	20.902

† Note: $\alpha = {}^{1}\!/_{\phi}$

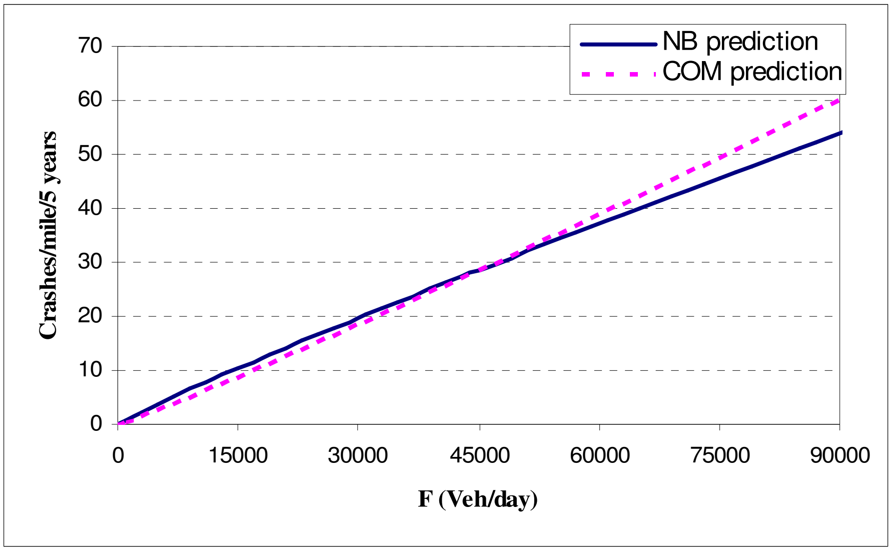


Figure 4.2: Estimated values (crashes/5-year) for the Texas data: NB and COM-Poisson models

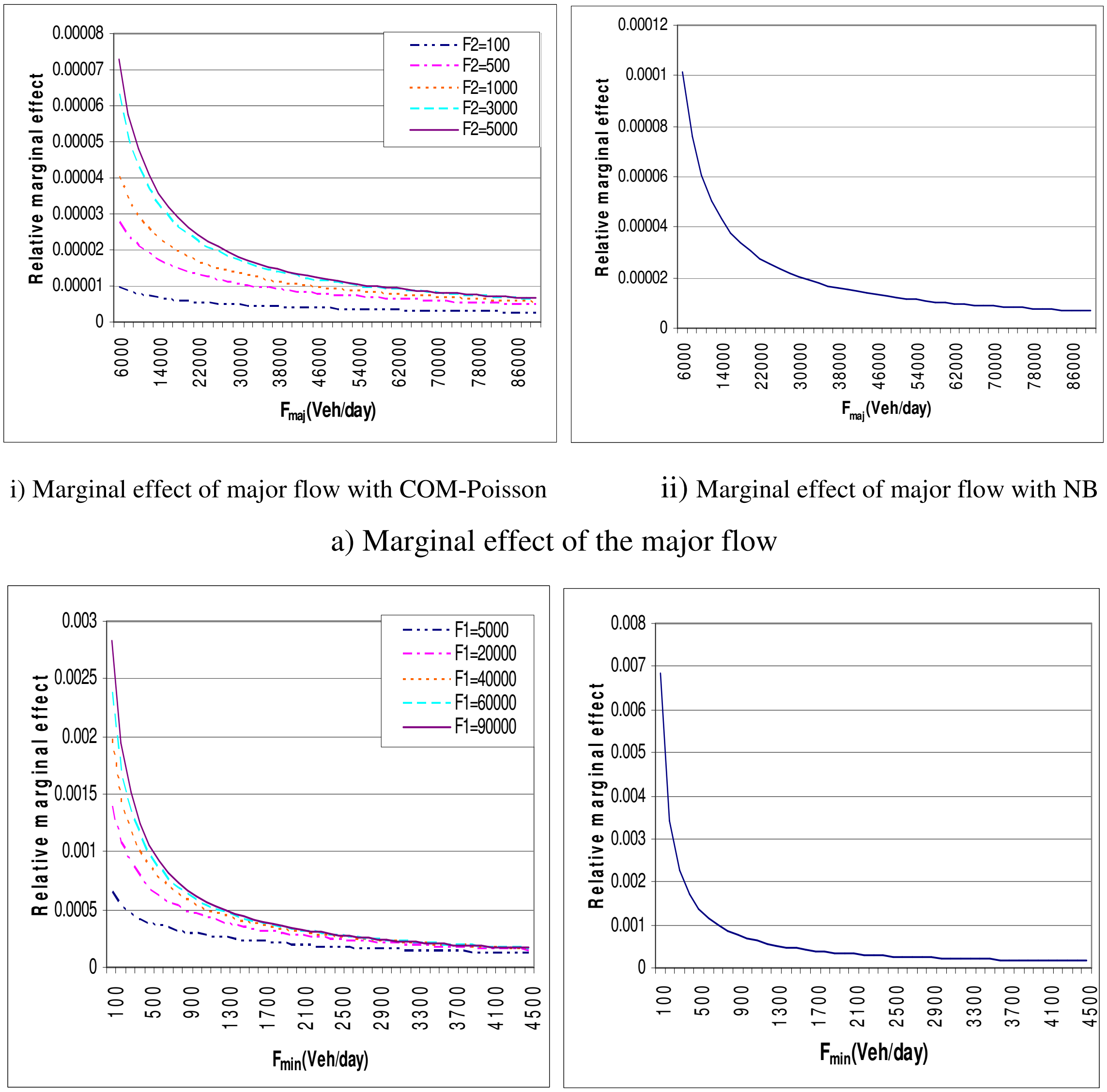
4.3.3 Marginal effects

An important issue in developing or using a regression model is the interpretation of the coefficients. Computing the marginal value of a particular variable can provide valuable information about how the regression coefficient related to that variable influence the expected mean value. For this exercise, the calculations of the marginal effect are slightly more complicated for the COM-Poisson distribution than for the NB distribution, which is usually straightforward. This is attributed to the fact that the parameter μ for the COM-Poisson is a centering parameter, as opposed to the expected mean value typically found in NB models.

The relative marginal effect of a particular variable or covariate X_i can be estimated using the following equation (Cameron and Trivedi, 1998): $\frac{1}{E[Y|X_i]}\frac{\partial E[Y|X_i]}{\partial X_i}$. In estimating the relative marginal effect of the variables, the mean approximation $E[Y] \approx \mu + 1/2\nu - 1/2$ can be used for the COM-Poisson models.

As seen in Figure 4.3a, the relative marginal effect of the major flow for the COM-Poisson model depends on the major and minor entering flows. This figure shows that the relative marginal effect on the expected mean for a unit increase in major flow remains nearly constant for lower minor flow volumes (e.g., 100 veh/day) and the rate of curvature increases with the increase in minor flows. On the other hand, the relative marginal effect of the major flow for the NB model is only independent of the minor flow. With a unit increase in major flow, the relative marginal effect on the expected mean value decreases. This decrease is smaller for higher major flows. Similar results can be seen for both COM-Poisson and NB models for the marginal effect related to the minor flow (Figure 4.3b).

Figure 4.4 shows that the relative marginal effect on the expected mean value decreases with a unit increase in flow for both COM-Poisson and NB models. In this figure, the y-axis is formatted under the logarithmic scale. In Figure 4.4, it can be seen that the marginal effect of traffic flow is higher for the NB model than for the COM-Poisson GLM. For the NB model, there is a sharp decrease in the marginal effect at lower flows and the curve decreases slightly for flows above 10,000 vehicles per day.



i) Marginal effect of major flow with COM-Poisson

ii) Marginal effect of major flow with NB

a) Marginal effect of the major flow

i) Marginal effect of minor flow with COM-Poisson

ii) Marginal effect of minor flow with NB

b) Marginal effect of the minor flow

Figure 4.3: Marginal effect of the traffic flows for the Toronto model

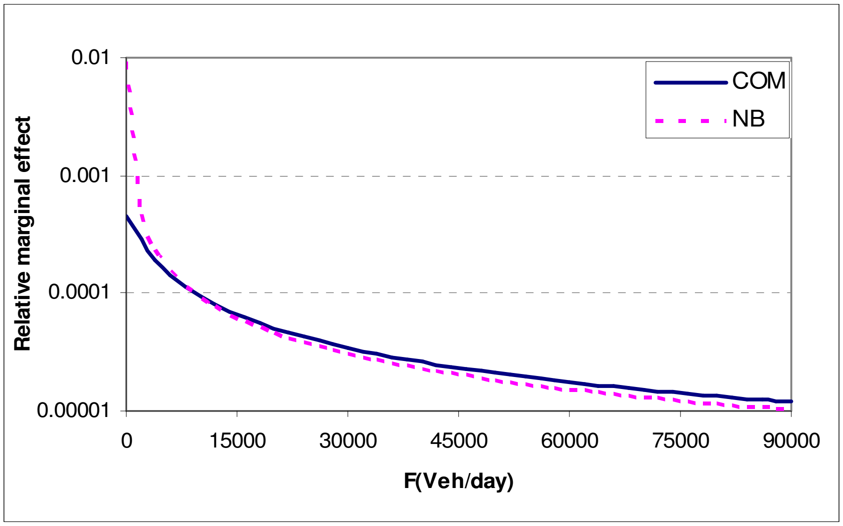


Figure 4.4: Marginal effect of traffic flow for the Texas model
(Note: y-axis is formatted under a logarithmic scale)

4.4 Discussion

This chapter has shown that the COM-Poisson GLM offers potential for modeling motor vehicle crashes. The following results are observed in this chapter:

- First, the model performs as well as the NB model (FB and MLE) for the functional form that only includes traffic flow as covariates. As detailed in the modeling results, both models provided similar GOF statistics and predictive performance. Guikema and Coffelt (2008) have reported similar comparison results between the COM-Poisson GLM and the FB NB model. The models used in Guikema and Coffelt (2008) included six covariates in both the centering and shape links. Hence, it is expected that COM-Poisson GLMs developed with several covariates, such as lane and shoulder widths, should work as well as the NB model. The performance of COM-Poisson with various covariates is shown in Chapter V.

- Second, although almost all crash datasets have been shown to exhibit over-dispersion (see Lord et al., 2005b), it has been documented that some crash datasets can sometimes experience under-dispersion (Oh et al., 2006). The NB GLM could theoretically handle under-dispersion, since the dispersion parameter can be negative ($Var\left(Y\right)=\mu+(-\alpha)\mu^{2}$). However, in this case, the mean of the Poisson is no longer gamma distributed because this latter distribution cannot have negative parameters (i.e., $gamma\left(\phi,\phi\right)$). In addition, researchers who have worked on the characterization of the NB distribution and GLM have indicated that a negative dispersion parameter could lead to a mis-specification of the PDF (when $-1/\left(\text{max of counts}\right)<\alpha$) (Clark and Perry, 1989; Saha and Paul, 2005). On the other hand, the COM-Poisson distribution has been shown to easily handle such datasets (Shmueli et al., 2005; Kadane et al., 2006; Geedipally et al., 2008; Guikema and Coffelt 2008; Lord et al., 2008c). The fact that the model handles under-dispersed data makes it more useful than the NB model, which has difficulty coping with this kind of data (as described above). Although not a good analysis approach, a transportation safety analyst could theoretically not have to worry about the characteristics of the dispersion in the data, since the COM-Poisson GLM can handle both over- and under-dispersed data, and a combination of both, if the data are characterized as such. The results about COM-Poisson handling an under-dispersed crash data is presented in the next chapter.

- Third, as discussed above, crash data can sometimes be subjected to very low sample mean values, which create data characterized by a large number of zeros (with the hypothesis that the space and time scales have been appropriately used, see Lord et al., 2005b). Consequently, NB models do not perform well with such datasets since they may tend to under-predict zero values (or over-estimate non-zero count values). To overcome this problem, some researchers have suggested the use of zero-inflated Poisson and NB models (Shankar et al., 1997). However, these models have been shown to be inappropriate for modeling crash data, since this kind of data does not exhibit two distinct generating processes, one of which is characterized by having a long-term mean equal to zero (which is not feasible for crash data) (Lord et al., 2005b; Warton, 2005; Wedagama et al., 2006).

Depending upon the specification of the parameters λ and ν, the COM-Poisson model can predict more zeros than the NB model for the same mean value. However, both models should not be used as a direct substitute to zero-inflated models (when they are warranted) (see Kadane et al., 2006). The performance of COM-Poisson distribution with the data characterized by SSS and LSM is shown in Chapter VI.

- Fourth, the COM-Poisson model is not significantly more difficult to implement than the FB NB model once the code for the maximum likelihood estimation is available. Sellers and Shmueli (2008) developed the code for MLE by the time this dissertation was written. Guikema and Coffelt (2008) developed the code needed to implement the COM-Poisson GLM in WinBUGS. For the models produced in this work, non-informative or vague priors were used for the regression coefficients. For the β coefficients, Normal (0,100) priors were used, for *ν,* a gamma (0.03, 0.1) prior was used, and for ϕ a gamma (0.1, 0.1) prior was used. The experimentation with other non-informative priors showed that the priors did not significantly affect either GOF of the models or the posterior parameter estimates. In addition, the difference in computational times for these models was not enormous. For example, for the full Toronto data set, a run of the COM-Poisson model with 35,000 replications took about 5 hours while a run of the FB NB model with 35,000 replications took between 1 and 1.5 hours; the absolute difference seems large, but some simulation runs can sometimes take up to two or three days in WinBUGS to converge depending on the complexity of the model hierarchical structure. Overall, implementing the COM-Poisson model is not significantly more difficult than implementing the FB NB model once the code for the COM-Poisson model is available.

4.5 Summary

The COM-Poisson distribution has the capability of handling the under-dispersed and over-dispersed count data. Crash data often exhibits over-dispersion. This chapter has documented the application of the COM-Poisson GLM for analyzing motor vehicle crash data exhibiting over-dispersion. The comparison between the COM-Poisson GLM with the NB model commonly used for analyzing motor vehicle crashes was also presented. The comparison analysis was carried out using the most common functional forms used by transportation safety analysts, which link crashes to the entering flows at intersections or on segments. Several methods were used to assess the statistical fit and predictive performance of the models.

The first section gave a brief methodology about the functional form used for modeling the crash mean and goodness-of-fit statistics used in the comparative analysis. The second section gave the description of the data used in this study. There were two datasets used in developing and comparing models. The first dataset contained crash data collected at 4-legged signalized intersections in Toronto, Ont. The second dataset included data collected for rural 4-lane divided and undivided highways in Texas. The third section gave the results of this study. The results showed that COM-Poisson GLMs perform as well as FB NB models in terms of GOF statistics and predictive performance. This result is supported by another recent study on this topic (Guikema and Coffelt, 2008). The estimated values with COM-Poisson and NB models are slightly different, with the COM-Poisson output being always lower than the NB output. The relative marginal effect of the major flow for the COM-Poisson model depends on the major and minor entering flows. The relative marginal effect on the expected mean for a unit increase in major flow remains nearly constant for lower minor flow volumes (e.g., 100 veh/day) and the rate of curvature increases with the increase in minor flows. On the other hand, the relative marginal effect of the major flow for the NB model is only independent of the minor flow. With a unit increase in major flow, the relative marginal effect on the expected mean value decreases. This decrease is smaller for higher major flows. Similar results can be seen for both COM-Poisson and NB models for the marginal effect related to the minor flow. The last section presented a brief discussion of the results. The next chapter describes the application of COM-Poisson GLM to crash data exhibiting under-dispersion.

CHAPTER V

APPLYING THE COM-POISSON GLM TO CRASH DATA EXHIBITING UNDER-DISPERSION

From Chapter IV, the results show that the COM-Poisson models have capability of handling crash data exhibiting over-dispersion. Although very rare, there is a possibility for the traffic crash data to exhibit under-dispersion when they are used in a context of generalized linear model (Oh et al., 2006; Park and Lord, 2007) and this phenomenon is less convenient to model (Oh et al., 2006). Many studies have demonstrated that the Poisson and NB models have significant difficulties handling (or cannot handle) data characterized by under-dispersion (Clark and Perry, 1989; Saha and Paul, 2005). The results of the Chapter III show that the COM-Poisson distribution is capable of handling under-dispersed data easily. Only a handful of studies have applied the COM-Poisson distribution and GLM to observed or simulated data characterized by under-dispersion (Kadane et al., 2005; Sellers and Shmueli, 2008; Guikema and Coffelt, 2008; Geedipally et al., 2008).

The objective of this chapter is to evaluate the performance of the COM-Poisson GLM for analyzing crash data exhibiting under-dispersion, in cases where Poisson and Poisson-gamma models cannot be used. To accomplish the objective of this study, several COM-Poisson models were estimated using crash data collected at 162 railway-highway crossings (RHX) in South Korea between 1998 and 2002. This dataset has been identified as being characterized by under-dispersion when the observations were modeled using regression methods (Oh et al., 2006). To model such dataset, Oh et al. (2006) have proposed the gamma probability model. The results will show that the COM-Poisson GLM can handle crash data when the modeling output shows signs of under-dispersion. The results also show that the model provides better statistical performance than the gamma probability and the traditional Poisson model.

The chapter is organized as follows: The first section describes the methodology used for estimating and comparing the various models. The second section presents the characteristics of the data used in this chapter. The third section summarizes the results of the parameter estimates and comparison analysis. The fourth section provides a useful discussion about the results. The last section gives the summary of this chapter.

5.1 Methodology

This section briefly describes the methodology used for comparing the different models. The same functional form used by Oh et al. (2006) was utilized for fitting all the models:

$$\mu_i = \exp(\beta_0 + \beta_1 \ln(F_i) + \sum_{j=1}^{n} \beta_j x_j) \tag{5.1}$$

where,

μ_i = the mean number of crashes for site i;

F_i = average daily vehicle traffic on site i (vehicles/day);

x_j = estimated covariates such as average daily railway traffic, detector distance etc; and,

β_i = estimated regression coefficients.

Different methods were used for evaluating the goodness-of-fit (GOF) and predictive performance of the models. The methods used in this study include the Deviance Information Criterion, Mean Absolute Deviance and Mean Squared Predictive Error. A brief explanation of these methods is given in Chapter IV.

The coefficients of the COM-Poisson GLMs were estimated using the software WinBUGS (Spiegelhalter et al., 2003). Vague or non-informative hyper-priors were utilized for the COM-Poisson GLMs. A total of 3 Markov chains were used in the model estimation process with 5,000 iterations per chain and the thinning parameter was set to 1. The first 2,500 iterations (burn-in samples) were discarded. Thus, the remaining 2,500 iterations were used for estimating the coefficients. The Gelman-Rubin (G-R) convergence statistic was used to verify that the simulation runs converged properly. In this analysis, the G-R statistic fell below 1.1 for all model parameters.

5.2 Data Description

This section provides an overview of the characteristics of the dataset used in this study. This dataset was previously used to develop Poisson and gamma probability models by Oh et al. (2006). The characteristics of the dataset used in this study are described in Table 5.1. It should be noted that looking at the raw observations, the crash data exhibit over-dispersion (mean=0.33, variance=0.36). The under-dispersion is in fact noticed when the observed values are modeled conditional on the mean, as described in the next section.

Table 5.1: Summary statistics of the dataset (Oh et al., 2006)

Variables		Min.	Max.	Average (std. dev)	Frequency
Crashes		0	3	0.33 (0.6)	162
AADT		10	61199	4617 (10391.57)	162
Average daily railway traffic		32	203	70.29 (37.34)	162
Presence of commercial area	1 (yes)	--	--	--	149 (91.98%)
	0 (no)	--	--	--	13 (8.02%)
Train detector distance		0	1329	824.5 (328.38)	162
Time duration between the activation of warning signals and gates		0	232	25.46 (25.71)	162
Presence of speed hump	1 (yes)	--	--	--	134 (82.72%)
	0 (no)	--	--	--	28 (17.28%)
Presence of track circuit controller	1 (yes)	--	--	--	113 (69.75%)
	0 (no)	--	--	--	49 (30.25%)
Presence of guide	1 (yes)	--	--	--	126 (77.78%)
	0 (no)	--	--	--	36 (22.22%)

Figure 5.1 below gives the comparison of the actual crash data distribution with the predicted values by Poisson distribution. The Poisson model predicts slightly lesser number of sites with zero and two crashes but more number of sites with one crash when compared to that of actual data.

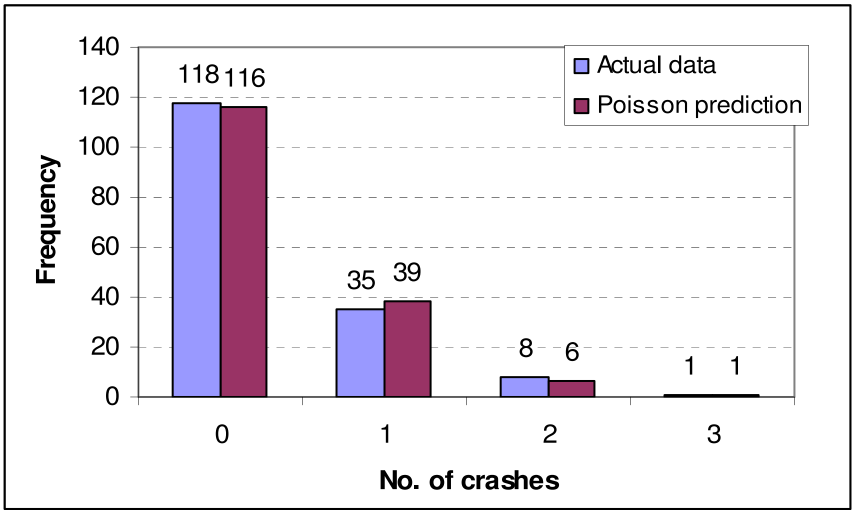
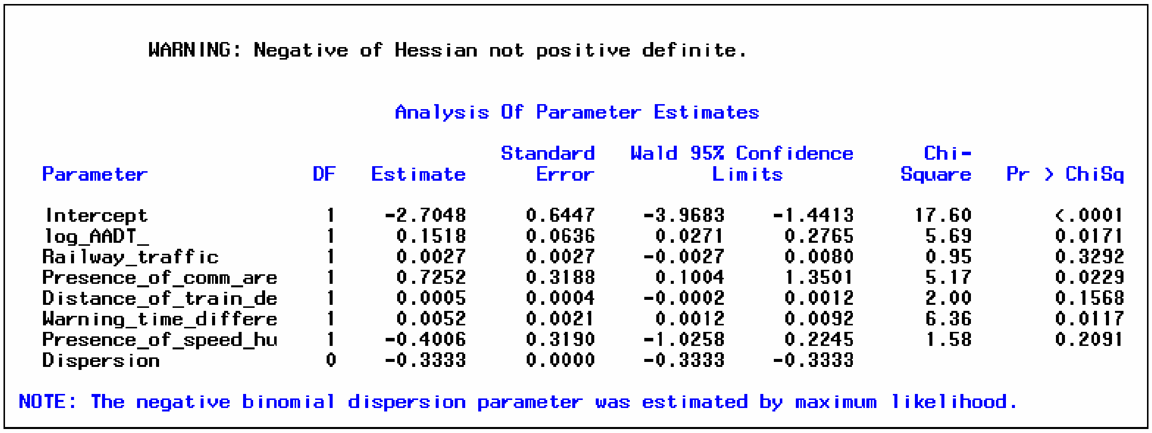


Figure 5.1: Observed crash data versus values estimated using the Poisson distribution

5.3 Results

This section describes the results of the analysis. Several models were estimated using the variables documented in Oh et al. (2006). They are described in Table 5.1. To evaluate the characteristics of the variance function, a Poisson-gamma model was first estimated using the six variables that were reported to be significant by the gamma probability model in the original study (Oh et al., 2006). Figures 5.2 and 5.3 show the output of the Poisson-gamma models for the MLE and Bayesian estimating methods, respectively. For the MLE, Figure 5.2 illustrates that the Poisson-gamma model cannot handle the data very well, as determined by the negative value of the dispersion parameter and its confidence interval (e.g., $Var(Y) = \mu + \alpha\mu^2$, where $\alpha =$ the dispersion parameter of the Poisson-gamma model). In addition, the model provides unreliable parameter estimates, since all the variables are not significant at the 5% level. For the Bayesian model, Figure

5.3 shows that the inverse dispersion parameter of the Poisson-gamma model becomes unstable and tends towards infinity (i.e., converges to a Poisson model), both when vague ($\phi \sim gamma(0.01,0.01)$) and non-vague hyper-priors ($\phi \sim gamma(0.2,0.1)$) are used. Similar to the MLE, most of the parameter estimates were not significant at the 5% level. In sum, these plots confirm that the modeling results are characterized by under-dispersion when the model is estimated using the six original explanatory variables or covariates.



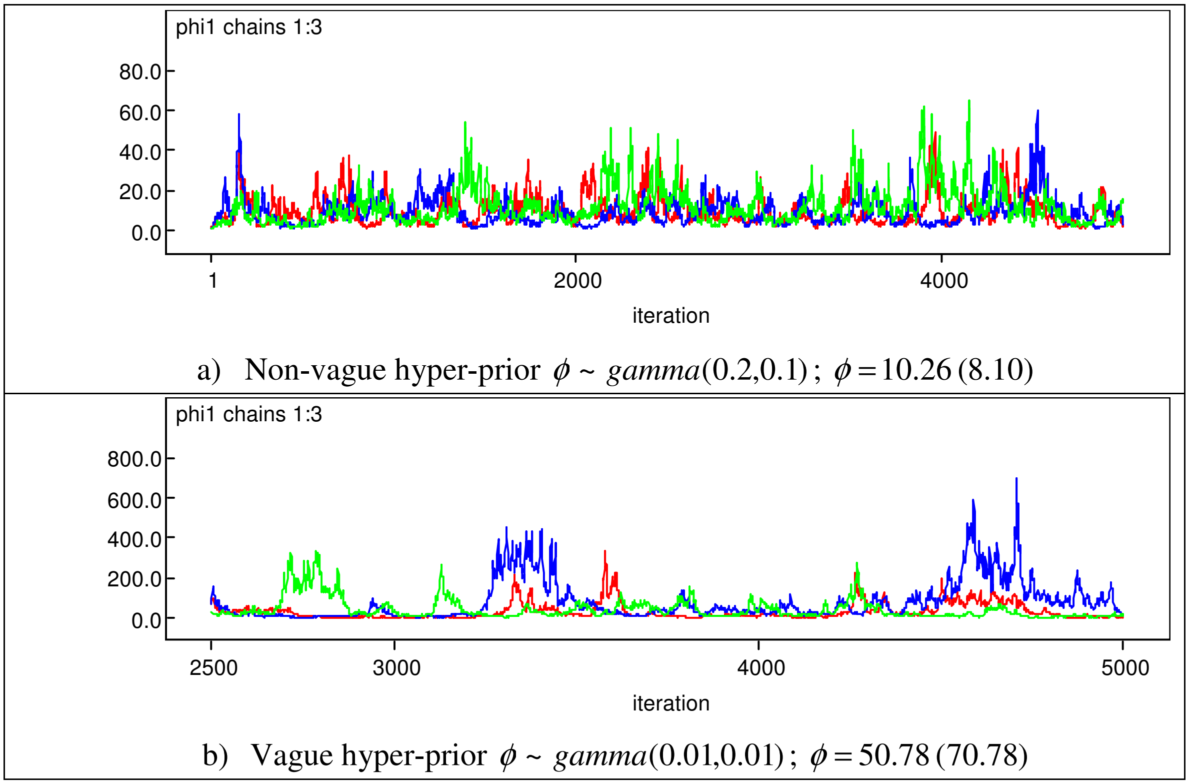
```
          WARNING: Negative of Hessian not positive definite.

                              Analysis Of Parameter Estimates

                                          Standard   Wald 95% Confidence    Chi-
Parameter              DF   Estimate       Error          Limits          Square   Pr > ChiSq

Intercept               1    -2.7048      0.6447    -3.9683    -1.4413     17.60       <.0001
log_AADT_               1     0.1518      0.0636     0.0271     0.2765      5.69       0.0171
Railway_traffic         1     0.0027      0.0027    -0.0027     0.0080      0.95       0.3292
Presence_of_comm_are    1     0.7252      0.3188     0.1004     1.3501      5.17       0.0229
Distance_of_train_de    1     0.0005      0.0004    -0.0002     0.0012      2.00       0.1568
Warning_time_differe    1     0.0052      0.0021     0.0012     0.0092      6.36       0.0117
Presence_of_speed_hu    1    -0.4006      0.3190    -1.0258     0.2245      1.58       0.2091
Dispersion              0    -0.3333      0.0000    -0.3333    -0.3333
NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.
```

Figure 5.2: SAS Output of the Poisson-gamma model



a) Non-vague hyper-prior $\phi \sim \mathit{gamma}(0.2,0.1)$; $\phi = 10.26\,(8.10)$

b) Vague hyper-prior $\phi \sim \mathit{gamma}(0.01,0.01)$; $\phi = 50.78\,(70.78)$

Figure 5.3: WinBUGS output (history for inverse dispersion parameter) of the Poisson-gamma model

As an initial step, a COM-Poisson model was developed with all eight explanatory variables documented in Table 5.1. In the subsequent step, individual models were developed by only considering the variables that were found to be significant for the Poisson and gamma probability models, respectively (all the variables are not the same). As seen in Table 5.2, Model 1, which contains the largest number of variables, is not the best model since the DIC penalizes models with a large number of parameters. Using this criterion, Model 2 is considered to be the best amongst all the models.

Table 5.2: Initial model comparison using COM-Poisson GLMs

Model Type[a]	DIC with COM-Poisson model
Model 1	201.520
Model 2	197.592
Model 3	205.309

[a] Model 1: significant variables identified by the Poisson and gamma models in Oh et al. (2006).
Model 2: significant variables identified by the Poisson model in Oh et al. (2006).
Model 3: significant variables identified by the gamma probability model in Oh et al. (2006).

Table 5.3 contains the parameter estimates and their associated standard errors for the three COM-Poisson GLMs. By including all eight variables, the COM-Poisson model shows that the modeling output is actually Poisson distributed, i.e. equi-dispersed (conditional on the mean). On the other hand, the model output shows signs of under-dispersion when the six variables found significant for the Poisson model are used. All the variables that were found to be significant with Poisson model in Oh et al. (2006) were also found to be significant for the COM-Poisson GLM. Interestingly, the COM-Poisson model output shows signs of over-dispersion when the same six variables that were found to be significant for the gamma model are utilized. More detailed comparisons between the models are described later.

Table 5.3: Parameter estimates of COM-Poisson models

Variables	Model 1	Model 2[a]	Model 3[b]
Constant	-5.859 (2.252)[c]	-4.616 (1.395)[c]	-6.131 (1.79)[c]
ADT	0.4295 (0.152)	0.4037 (0.109)	0.3631 (0.145)
Average daily railway traffic	0.0028 (0.004)	--	0.0032 (0.006)
Presence of commercial area	0.9244 (0.459)	0.8966 (0.468)	1.108 (0.685)
Train detector distance	0.0016 (0.0009)	0.0012 (0.0006)	0.0014 (0.0008)
Time duration between the activation of warning signals and gates	0.0032 (0.004)	--	0.0057 (0.006)
Presence of track circuit controller	-0.8361 (0.444)	-0.9068 (0.433)	--
Presence of guide	-0.5869 (0.454)	-0.6414 (0.430)	--
Presence of speed hump	-0.9933 (0.514)	-1.039 (0.506)	-1.077 (0.732)
Shape Parameter (ν_0)	0.9578	1.177	0.7876
Deviance	193.8	192.0	200.6
DIC	201.520	197.592	205.309

[a] Based on the Poisson model documented in Oh et al. (2006).
[b] Based on the gamma probability model documented in Oh et al. (2006).
[c] Posterior standard error.

Table 5.4 summarizes the direct comparison between the Poisson model developed in Oh et al. (2006) and the COM-Poisson model. This table shows that the COM-Poisson model is found to provide a slightly better statistical fit than the traditional Poisson model using the DIC. As seen from the shape parameter of COM-Poisson distribution (ν_0), the model output exhibits under-dispersion. With the exception of the magnitude of the parameter estimates, all the variables in both models have similar effect on the predicted values. As seen in Table 5.4, the GOFs are calculated using the mean and the centering parameters of Shmueli et al. (2005) and Guikema and Coffelt (2008), respectively.

Table 5.4: Comparison of parameter estimates between the Poisson and the COM-Poisson model

Variables	**Poisson Estimates**[a]	**COM-Poisson Estimates (Model 2)**	
Constant	-5.406	-4.616	
ADT	0.460	0.4037	
Presence of commercial area	0.975	0.8966	
Train detector distance	0.0016	0.0012	
Presence of track circuit controller	-0.917	-0.9068	
Presence of guide	-0.613	-0.6414	
Presence of speed hump	-1.063	-1.039	
Shape Parameter (ν_0)	--	1.177	
Deviance	203.5	192.0	
DIC	211.444	197.592	
MAD	0.354	Using μ	0.361
		Using λ	0.346
		Using E[Y]	0.312
MSPE	0.243	Using μ	0.241
		Using λ	0.254
		Using E[Y]	0.253

[a] Parameter estimates are directly taken from Oh et al. (2006); no standard errors were provided.

Figure 5.4 compares the estimated values of the Poisson and COM-Poisson models with the change in traffic flow. These values were estimated using the average value for the train detector distance, and the absence of commercial area, track circuit controller, guide and speed hump. An important point to note is that the mean of the COM-Poisson GLM ($E[Y]$) was not used for estimating the number of crashes. This is attributed to the fact that the mean of the COM-Poisson distribution is always zero for μ<0.3. Since a long-term mean equal to zero is not feasible for crash data analysis (see Lord et al., 2005b & 2007), the centering parameter is used for estimating the number of crashes. As seen in Table 5.4, the model fit is not much different when the centering parameter or the posterior mean is utilized. Figure 5.4 shows that the predicted values are always lower for the COM-Poisson model than for the Poisson model.

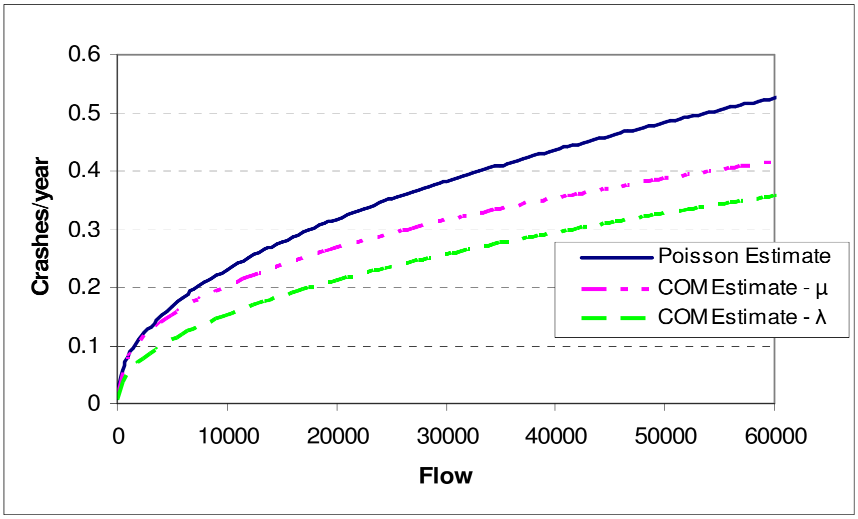


Figure 5.4: Estimated values from the Poisson and the COM-Poisson GLM

The direct comparison between the gamma probability and COM-Poisson models is summarized in Table 5.5. Out of the six variables that were found to be significant with the gamma probability model, only four variables were significant for the COM-Poisson GLM. Surprisingly, the COM-Poisson model shows that the modeling output exhibits over-dispersion whereas the gamma probability model output shows under-dispersion. This disparity is explained by the differences in the mean values between both models. Since the characteristics of the dispersion are conditional on the mean, different mean values (estimated by the model) will lead to different degrees of dispersion. The GOF statistics show that the COM-Poisson model provides better statistical fit than the gamma probability model. By visually examining the rate of change, it can be observed that the marginal effect of each variable is larger for the COM-Poisson model than that of the gamma probability model.

Table 5.5: Comparison of parameter estimates between the gamma probability and the COM-Poisson model

Variables	Gamma Estimates[a]	COM-Poisson Estimates (Model 3)	
Constant	-3.438	-6.131	
ADT	0.230	0.3631	
Average daily railway traffic	0.004	0.0032[b]	
Presence of commercial area	0.651	1.108	
Train detector distance	0.001	0.0014	
Time duration between the activation of warning signals and gates	0.004	0.0057[b]	
Presence of speed hump	-1.58	-1.077	
Shape Parameter	2.062	0.788	
MAD	0.459	Using μ	0.331
		Using λ	0.355
		Using E[Y]	0.325
MSPE	0.308	Using μ	0.305
		Using λ	0.287
		Using E[Y]	0.301

[a] Parameter estimates are directly taken from Oh et al. (2006); no standard errors were provided.

[b] Not found to be significant at the 5% confidence level.

As discussed above, the centering parameters (λ and μ) were used for the COM-Poisson GLM for estimating the number of crashes. The estimated values produced by the gamma probability model and COM-Poisson models are compared with the change in traffic flow and are shown in Figure 5.5. These values are estimated using the average value for the train detector distance, daily railway traffic, warning time duration, and the absence of commercial area, and speed hump. Figure 5.5 illustrates that the estimate μ is very similar to the mean values estimated by the gamma probability model.

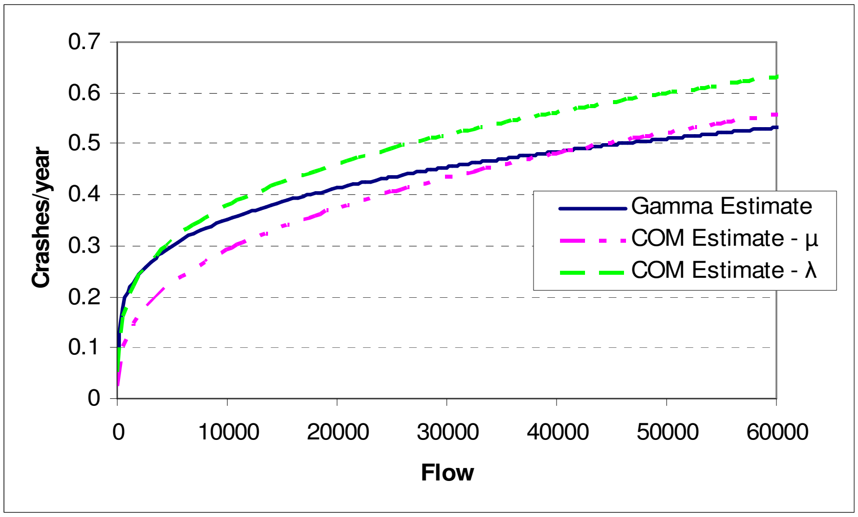


Figure 5.5: Estimated values for the gamma and the COM-Poisson GLM

5.4 Discussion

The results of this study show that the COM-Poisson GLM is quite flexible for handling crash data exhibiting under-dispersion (when conditional on the mean). They support other studies on this topic (Guikema and Coffelt, 2008; Geedipally et al., 2008; Sellers and Shmueli, 2008), which shows that COM-Poisson GLM can both handle over- and under-dispersion. Given this outcome, the results of the analysis presented above still raise a few important topics that merit further discussion.

- First, the COM-Poisson model performs well for the typical functional forms that have been used by traffic safety analysts (e.g., $\mu = \beta_0 F^{\beta_1} e^{\sum x_i \beta_i}$). This study has shown that the COM-Poisson models provide better statistical performance than the Poisson and gamma probability models when under-dispersion is observed in the data. Furthermore, as discussed above, the gamma probability model works as a dual-state model. Although this model can handle under- dispersion (and over-dispersion), it may not be appropriate for analyzing crash data (Lord et al., 2005b, 2007). Consequently, the COM-Poisson GLM offers a more defensible approach for modeling under-dispersed data, since it does not assume a dual-state data generating process.

- Second, with the inclusion of all eight significant variables, the COM-Poisson GLM shows that the modeling output exhibits near equi-dispersion. This characteristic has also been documented in Miaou and Song (2005), who proposed complex multivariate hierarchical predictive models for analyzing crash data. They showed that by significantly improving the mean function, one could almost eliminate the over-dispersion. In this study, when the two non-significant variables were removed, the COM-Poisson model clearly showed signs of over-dispersion. On the other hand, when the same six variables used in the Poisson model were utilized for the COM-Poisson model, the modeling output showed signs of under-dispersion. This unusual characteristic (i.e., changes from under-dispersion to over-dispersion or vice-versa by including or removing explanatory variables) is attributed to the very low sample mean value (0.33 crash per year) of the dataset. Under this condition, this illustrates that the predicted mean values can significantly influence the variation found in the data.

- Third, the COM-Poisson GLM is very sensitive to the selection of the model's parameters. During the modeling process, some simulation runs took up to two days for the MCMC replications to properly converge (as discussed above, the G-R Statistic fell below 1.1 for all parameters); hence, the number of iterations was limited to 5,000. The length of the simulation can be greatly influenced by the selection of the hyper-priors (location of the upper and lower boundaries). Thus, the analyst must be careful in defining the appropriate hyper-priors, especially when the sample mean value is very small (this was also discussed in Lord and Miranda-Moreno, 2008). It is anticipated, however, that the computational time will be decreased significantly when the likelihood function of the COM-Poisson distribution becomes available. Sellers and Shmueli (2008) are currently developing a likelihood formulation for this distribution. Despite the computational time advantage, the MLE does not provide the full posterior distributions for the regression parameters nor does it allow expert knowledge to be incorporated through the use of informative priors (see, e.g., Washington and Oh, 2006, and Miranda-Moreno et al., 2008, for additional information about how expert knowledge can improve the performance of Bayesian models in highway safety).

5.5 Summary

Although very rare, crash data sometimes exhibit under-dispersion. There are numerous reasons for data to show under-dispersion, one of which is the existence of large number of zeros, which in turn have a low sample mean value. The COM-Poisson distribution has the capability of handling under-dispersed data easily. The objective of this chapter was to evaluate the performance of the COM-Poisson GLM for analyzing crash data exhibiting under-dispersion (when conditional on the mean). The modeling results of the COM-Poisson were then compared to those produced from the Poisson and gamma probability models documented in Oh et al. (2006).

The first section gave a brief methodology about the functional form for modeling the crash mean and the testing protocol used in the study. The second section presented data description. Crash data collected at 162 railway-highway crossings in South Korea between 1998 and 2002 was used in this study. This dataset has been shown to exhibit under-dispersion when models linking crash data to various explanatory variables are estimated. The third section gave the results of the study. The results showed that the COM-Poisson GLM can handle crash data when the modeling output shows signs of under-dispersion. They also showed that the model analyzed in this study provides better statistical performance than the gamma probability and the traditional Poisson models, at least for this dataset. Similarly, the COM-Poisson GLM offers a more defensible approach than the gamma probability model, since the former does not assume that the observed data follow a dual-state generating process. Given the changes in the nature of the variance function when variables are included or excluded, it is possible that such datasets could contain intermingled over- and under-dispersed counts. The last section presented a brief summary of the results. The next chapter describes the effect of low sample mean and small sample size on the parameter estimates of the COM-Poisson distribution.

CHAPTER VI

EFFECTS OF SMALL SAMPLE SIZE AND LOW SAMPLE MEAN ON PARAMETER ESTIMATES OF THE COM-POISSON DISTRIBUTION

Crash data are often characterized by small sample size and low sample mean. It is a usual practice for traffic safety analysts to develop statistical models using the limited number of observations where data can be collected (Lord, 2000 and Oh et al., 2003). This small sample size problem is usually attributed to the prohibitive costs involved in collected the crash data and the variables influencing the crash occurrence (Lord and Bonneson, 2005). Due to the existence of large number of zeroes, crash data usually exhibit a distribution with a low sample mean.

There are numerous studies explaining the effects of data characterized by low sample mean on NB models in the traffic safety literature. The results concerning the goodness-of-fit, dispersion parameter, parameter estimates and confidence intervals may be biased for the NB models when the data are characterized by small sample size and low sample mean values. It is clear from the Chapter III that the COM-Poisson models do not perform well when the data are characterized by the low sample mean. It is important to determine potential bias in the estimation of the parameters of COM-Poisson models when the data are characterized by the low sample mean and small sample size.

The first objective of this research is to know whether the centering parameter (μ) and shape parameter (ν) are properly estimated when the data are characterized by LSM and SSS. The bias in the parameter estimation will be then calculated. Secondly, the influence of the assumption of various prior distributions on the shape parameter in the posterior estimation is evaluated. To do this, initially a log-normal distribution is assumed, and then followed by a gamma distribution for the shape parameter. The third objective consisted of determining the recommended minimum sample size for developing COM-Poisson models subjected to low sample mean values and small sample size. This recommendation is to reduce unreliable estimation of the posterior mean of the centering and shape parameter. To accomplish the objectives, a series of COM-Poisson distributions were simulated using different values describing the centering parameter, the shape parameter, and the sample size.

This chapter is organized as follows: First section gives a brief note on the simulation framework used in this study. The second section presents the simulation results with both the log-normal prior and gamma prior for shape parameter. It also gives the recommended sample size for a given sample mean. The third section gives a brief discussion about the results. The last section summarizes the chapter.

6.1 Simulation Framework

This section briefly describes the simulation study that illustrates the effects of LSM and SSS on the prediction of centering and shape parameter of COM-Poisson models. The centering parameter is simulated from the log-normal distribution and the shape parameter is simulated from log-normal distribution and gamma distribution for two different scenarios. The data are then simulated from the PDF of COM-Poisson distribution. The sample size and sample mean were chosen in such a way that they represented normal conditions and extreme conditions (SSS and/or LSM).

The following steps give brief overview of the simulation framework:

1. Generate a centering parameter (μ_i) for an observation i from a fixed value μ.

   $$\mu_i = \mu$$

2. Generate a shape parameter (ν_i) for an observation i from a fixed value ν.

   $$\nu_i = \nu$$

3. Select an appropriate relative error ε ($\varepsilon = 0.01$ is considered in this study)

4. Generate a discrete value (Y_i) for an observation i from a COM-Poisson distribution with centering parameter μ_i, shape parameter ν_i and relative error ε

   $Y_i \sim$ COM-Poisson (μ_i, ν_i, ε)
5. Repeat steps 1 to 4 for '*n*' number of times where '*n*' refers to the required sample size.

The simulation was carried out in the MATLAB® 7.1.0 R14 (The Mathworks Inc, Natick, MA). The following are the scenarios considered in the simulation study:

Expected sample mean $E\ (Y_i) \cong$ **0.5**, **1.0**, 10
Shape parameter (ν) = **0.4**, 0.6, 0.8
Sample size (n) = **50**, **100**, 1000

The number in the bold character represents the extreme values characterized by low sample mean and/or small sample size. Since the expected mean value cannot be directly given as an input in the COM-Poisson distribution for simulating discrete values, the centering and shape parameters were properly selected so as to generate the approximate sample mean value. Table 6.1 gives the centering and shape parameter values that were used to generate the above mentioned sample means:

The MCMC implementation of the COM GLM proposed by Guikema and Coffelt (2008) was used for the model estimation process. Non-informative log-normal priors (i.e., Log-N (0,100) priors) were utilized for the centering and shape parameters of COM GLMs in the first scenario. Non-informative log-normal prior (i.e., Log-N (0,100) prior) for the centering parameter and a non-informative gamma prior (i.e., gamma (0,100) prior) for the shape parameter were utilized in the second scenario. A total of 3 Markov chains with 50,000 iterations were used initially to check the convergence. A satisfactory convergence was achieved for 50,000 iterations. Then, a single chain with 100,000 iterations and a thinning of 10 were used in the model estimation process. The first 50,000 iterations (burn-in samples) were discarded. The remaining iterations were used for estimating the coefficients. For each combination of sample size, centering parameter, and shape parameter, the simulation was replicated 100 times. At the end of replications, the statistics such as minimum value, maximum value, mean and standard deviation were computed.

Table 6.1: Centering and shape parameters for a specified sample mean

Centering parameter	Shape parameter	Sample mean
$\mu = 10$	$\nu = 0.8$	$E\ (Y_i)\ \cong\ 10$
$\mu = 1.0$		$E\ (Y_i)\ \cong\ 1.0$
$\mu = 0.5$		$E\ (Y_i)\ \cong\ 0.5$
$\mu = 10$	$\nu = 0.6$	$E\ (Y_i)\ \cong\ 10$
$\mu = 0.8$		$E\ (Y_i)\ \cong\ 1.0$
$\mu = 0.3$		$E\ (Y_i)\ \cong\ 0.5$
$\mu = 10$	$\nu = 0.4$	$E\ (Y_i)\ \cong\ 10$
$\mu = 0.5$		$E\ (Y_i)\ \cong\ 1.0$
$\mu = 0.14$		$E\ (Y_i)\ \cong\ 0.5$

6.2 Simulation Results

This section summarizes the simulation results for both the scenarios. The first section summarizes the results for the assumption of log-normal prior for the shape parameter. The second section gives the results for the assumption of gamma prior for the shape parameter.

6.2.1 Log-normal prior

This section presents the results for the assumption of log-normal prior for the shape parameter during the posterior estimation.

To test the precision in the estimation of the parameters, the simulation runs were performed for a sample mean of 10 initially and the results are presented in Table 6.2 below. For the sample size of 1,000, all the parameters are accurately estimated and the theoretical value of each parameter is almost equal to its predicted value. For the sample size of 100 and 50, the theoretical value of each parameter is accurately predicted but the standard deviation in the prediction becomes larger when compared to sample size of 1,000. In other words, as the sample size decreased the standard deviation increased. The minimum and maximum values of each parameter became noticeable when the sample size decreased to 50.

Table 6.2: Results of parameters for $E[Y] \cong 10$ (log-normal prior)

	N=1000			N=1000			N=1000		
	$\mu = 10$	**$\nu = 0.8$**	**E[Y]=10.05**	**$\mu = 10$**	**$\nu = 0.6$**	**E[Y]=10.14**	**$\mu = 10$**	**$\nu = 0.4$**	**E[Y]=10.55**
Mean	9.8887	0.7748	9.9083	9.8031	0.6151	9.9331	9.8385	0.4211	10.3554
Std.dev	0.1050	0.0393	0.0989	0.1294	0.0280	0.1247	0.1755	0.0218	0.1468
Min.	9.5126	0.6650	9.5828	9.4022	0.5283	9.6156	9.3982	0.3631	9.9532
Max.	10.2622	0.8869	10.2896	10.1762	0.6776	10.2617	10.4199	0.4894	10.8348
	N=100			N=100			N=100		
Mean	9.8612	0.8015	9.8899	9.7646	0.6118	9.9311	9.7413	0.4234	10.2797
Std.dev	0.3690	0.1222	0.3663	0.4094	0.0890	0.3760	0.6023	0.0669	0.5200
Min.	8.9236	0.5290	8.9539	8.6515	0.4056	8.7789	8.1665	0.2733	8.8552
Max.	10.9703	1.1756	10.9538	10.8474	0.9242	10.9251	11.3726	0.6472	11.5918
	N=50			N=50			N=50		
Mean	9.8755	0.8080	9.9114	9.7141	0.6206	9.9235	9.6895	0.4207	10.2667
Std.dev	0.5425	0.1744	0.5187	0.7046	0.1630	0.5965	0.8512	0.0925	0.7066
Min.	8.3475	0.4695	8.5334	7.5712	0.2824	8.3772	7.1542	0.2494	8.3387
Max.	11.0571	1.4233	11.2087	11.5747	1.1557	11.6667	11.6500	0.7261	12.3337

The simulation results for the sample mean of 1 are presented in Table 6.3. The predicted values are slightly mis-estimated compared to the theoretical value for all the sample sizes. As the sample size decreased, the standard deviation increased. As expected, the value of standard deviation, minimum and maximum values of the posterior means became highly noticeable for the sample size of 50. The change in the shape parameter value did not influence the estimation of centering parameter and sample mean.

Table 6.3: Results of parameters for $E[Y] \cong 1$ (log-normal prior)

	N=1000			N=1000			N=1000		
	$\mu = 1$	**$\nu = 0.8$**	**E[Y]=1.09**	**$\mu = 0.8$**	**$\nu = 0.6$**	**E[Y]=1.07**	**$\mu = 0.5$**	**$\nu = 0.4$**	**E[Y]=1.04**
Mean	1.0100	0.8598	0.9780	0.8906	0.7562	0.9376	0.7195	0.6346	0.9034
Std.dev	0.0643	0.0576	0.0262	0.0724	0.0482	0.0283	0.0893	0.0636	0.0442
Min.	0.8220	0.7073	0.9095	0.6999	0.6300	0.8732	0.3558	0.3510	0.8169
Max.	1.2046	1.0756	1.0598	1.0485	0.8692	1.0005	0.9219	0.7703	1.0369
	N=100			N=100			N=100		
Mean	0.7883	0.7419	0.9482	0.6245	0.5857	0.9126	0.3971	0.4076	0.9498
Std.dev	0.3389	0.3164	0.1925	0.3652	0.3062	0.2655	0.3019	0.2439	0.2781
Min.	0.0000	0.0593	0.0000	0.0000	0.0624	0.0000	0.0000	0.0637	0.0000
Max.	1.4272	1.6961	1.5425	1.6149	1.6359	1.6719	1.1787	1.2473	1.7624
	N=50			N=50			N=50		
Mean	0.6320	0.6428	0.9139	0.4605	0.4686	0.9007	0.3039	0.3474	0.9116
Std.dev	0.4170	0.4038	0.2840	0.3894	0.3437	0.3483	0.3234	0.3017	0.3935
Min.	0.0000	0.0509	0.0000	0.0000	0.0496	0.0000	0.0000	0.0572	0.0000
Max.	1.5574	1.7742	1.5168	1.4384	1.6644	1.5276	1.5007	1.4116	1.5797

Table 6.4 presents the simulation results for the sample mean value equal to 0.5. This table exhibits similar characteristics as those shown in Table 6.3. For the sample mean equal to 0.5, the estimators are highly unreliable. For the sample size of 50, most of the estimated values are not significantly different from zero. Also the minimum value for most of the estimators is very small and is almost equal to zero.

Table 6.5 presents the results of bias in the prediction of the parameters. The bias of the centering parameter μ and shape parameter ν is calculated as the difference between their expected value and the theoretical value. This table shows that the bias in parameter estimation follows no specific trend with the change in sample size or sample mean value.

Table 6.4: Results of parameters for $E[Y] \cong 0.5$ (log-normal prior)

	N=1000			N=1000			N=1000		
	$\mu = 0.5$	**$\nu = 0.8$**	**E[Y]=0.53**	**$\mu = 0.3$**	**$\nu = 0.6$**	**E[Y]=0.47**	**$\mu = 0.14$**	**$\nu = 0.4$**	**E[Y]=0.46**
Mean	0.7173	1.5330	0.3504	0.5871	1.2898	0.3145	0.5364	1.1990	0.3021
Std.dev	0.0502	0.1019	0.0192	0.0504	0.0899	0.0154	0.0552	0.0925	0.0129
Min.	0.5819	1.2652	0.3090	0.4067	0.9851	0.2736	0.3666	0.9323	0.2653
Max.	0.8446	1.8036	0.4762	0.7167	1.5424	0.4448	0.6600	1.4187	0.3311
	N=100			N=100			N=100		
Mean	0.4489	1.0212	0.4026	0.3073	0.7950	0.3463	0.2505	0.6685	0.3655
Std.dev	0.2670	0.5555	0.1375	0.2424	0.5336	0.1654	0.2278	0.4903	0.1805
Min.	0.0000	0.0951	0.0000	0.0000	0.1003	0.0000	0.0000	0.1055	0.0000
Max.	0.9468	2.4268	0.9824	0.9062	2.3888	0.8273	0.7826	2.5927	0.7655
	N=50			N=50			N=50		
Mean	0.2901	0.7927	0.3892	0.2179	0.7071	0.3496	0.1756	0.5544	0.3441
Std.dev	0.2610	0.7706	0.1813	0.2353	1.1507	0.2115	0.2217	0.6194	0.2074
Min.	0.0000	0.1016	0.0000	0.0000	0.1016	0.0000	0.0000	0.0960	0.0000
Max.	1.0927	6.9424	0.8533	0.9691	14.5515	0.8145	0.9046	5.2199	0.7849

Table 6.5: Bias in the parameter estimation (log-normal prior)

$E[Y] \cong 10$		**$\mu = 10$**	**$\nu = 0.8$**	**E[Y] = 10.05**	**$\mu = 10$**	**$\nu = 0.6$**	**E[Y] = 10.14**	**$\mu = 10$**	**$\nu = 0.4$**	**E[Y] = 10.55**
	N=1000	0.1113	0.0252	0.1417	0.1969	-0.0151	0.2069	0.1615	-0.0211	0.1946
	N=100	0.1388	-0.0015	0.1601	0.2354	-0.0118	0.2089	0.2587	-0.0234	0.2703
	N=50	0.1245	-0.0080	0.1386	0.2859	-0.0206	0.2165	0.3105	-0.0207	0.2833
$E[Y] \cong 1$		**$\mu = 1$**	**$\nu = 0.8$**	**E[Y] = 1.09**	**$\mu = 0.8$**	**$\nu = 0.6$**	**E[Y] = 1.07**	**$\mu = 0.5$**	**$\nu = 0.4$**	**E[Y] = 1.04**
	N=1000	-0.0100	-0.0598	0.1120	-0.0906	-0.1562	0.1324	-0.2195	-0.2346	0.1366
	N=100	0.2117	0.0581	0.1418	0.1755	0.0143	0.1574	0.1029	-0.0076	0.0902
	N=50	0.3680	0.1572	0.1761	0.3395	0.1314	0.1693	0.1961	0.0526	0.1284
$E[Y] \cong 0.5$		**$\mu = 0.5$**	**$\nu = 0.8$**	**E[Y] = 0.53**	**$\mu = 0.3$**	**$\nu = 0.6$**	**E[Y] = 0.47**	**$\mu =$ 0.14**	**$\nu = 0.4$**	**E[Y] = 0.46**
	N=1000	-0.2173	-0.7330	0.1796	-0.2871	-0.6898	0.1555	-0.3964	-0.7990	0.1579
	N=100	0.0511	-0.2212	0.1274	-0.0073	-0.1950	0.1237	-0.1105	-0.2685	0.0945
	N=50	0.2099	0.0073	0.1409	0.0821	-0.1071	0.1204	-0.0356	-0.1544	0.1159

The mean squared error (MSE) of an estimator quantifies the amount by which an estimator differs from the theoretical value of the estimator being estimated. The MSE of the estimated centering parameter $\hat{\mu}$ with respect to its theoretical value μ is defined as

$$\text{MSE}(\hat{\mu}) = \text{Var}(\hat{\mu}) + (\text{Bias}(\hat{\mu}, \mu))^2$$

The bias of centering parameter 'μ' is calculated as $E(\hat{\mu}) - \mu$

Similarly the MSE of the shape parameter estimator $\hat{\nu}$ is defined as

$$\text{MSE}\ (\hat{\nu}) = \text{Var}\ (\hat{\nu}) + (\text{Bias}\ (\hat{\nu},\ \nu))^2$$

The bias of shape parameter 'ν' is calculated as $E(\hat{\nu}) - \nu$

As seen from Table 6.6, the MSE increases as the sample size and sample mean decreases. There is always a systematic bias in the prediction of sample mean as the sample size decreased. This is not always true in case of centering and shape parameter.

Table 6.6: MSE of the estimated parameters (log-normal prior)

$E[Y] \cong 10$		**$\mu = 10$**	**$\nu = 0.8$**	**E[Y] = 10.05**	**$\mu = 10$**	**$\nu = 0.6$**	**E[Y] = 10.14**	**$\mu = 10$**	**$\nu = 0.4$**	**E[Y] = 10.55**
	N=1000	0.0234	0.0022	0.0299	0.0555	0.0010	0.0584	0.0569	0.0009	0.0594
	N=100	0.1554	0.0149	0.1598	0.2230	0.0081	0.1851	0.4297	0.0050	0.3434
	N=50	0.3098	0.0305	0.2883	0.5782	0.0270	0.4027	0.8210	0.0090	0.5795
$E[Y] \cong 1$		**$\mu = 1$**	**$\nu = 0.8$**	**E[Y] = 1.09**	**$\mu = 0.8$**	**$\nu = 0.6$**	**E[Y] = 1.07**	**$\mu = 0.5$**	**$\nu = 0.4$**	**E[Y] = 1.04**
	N=1000	0.0042	0.0069	0.0132	0.0135	0.0267	0.0183	0.0561	0.0591	0.0206
	N=100	0.1597	0.1035	0.0572	0.1641	0.0940	0.0953	0.1018	0.0596	0.0855
	N=50	0.3094	0.1877	0.1117	0.2669	0.1354	0.1500	0.1430	0.0938	0.1713
$E[Y] \cong 0.5$		**$\mu = 0.5$**	**$\nu = 0.8$**	**E[Y] = 0.53**	**$\mu = 0.3$**	**$\nu = 0.6$**	**E[Y] = 0.47**	**$\mu =$ 0.14**	**$\nu = 0.4$**	**E[Y] = 0.46**
	N=1000	0.0497	0.5477	0.0326	0.0850	0.4839	0.0244	0.1602	0.6469	0.0251
	N=100	0.0739	0.3575	0.0351	0.0588	0.3227	0.0426	0.0641	0.3125	0.0415
	N=50	0.1122	0.5939	0.0527	0.0621	1.3355	0.0592	0.0504	0.4075	0.0564

Table 6.7 presents the values for the cutoff factor. This factor is used to know the approximate sample size for the given mean. Generally a threshold level of 85% or 90% is considered in most of the cases. The threshold level of 85% was considered in this research. The values above the given threshold level are retained. If the value falls below the threshold value then the sample size is increased for the given mean until the factor reaches the target significant level. The cutoff factor is defined as:

$$\text{Cutoff factor} = \frac{Mean - \sqrt{MSE}}{Mean}$$

Table 6.7: Cutoff factor of the estimated parameters (log-normal prior)

		$\mu = 10$	$\nu = 0.8$	**E[Y] = 10.05**	$\mu = 10$	$\nu = 0.6$	**E[Y] = 10.14**	$\mu = 10$	$\nu = 0.4$	**E[Y] = 10.55**
$E[Y] \cong 10$	N=1000	0.9845	0.9397	0.9826	0.9760	0.9483	0.9757	0.9758	0.9278	0.9765
	N=100	0.9600	0.8476	0.9596	0.9516	0.8532	0.9567	0.9327	0.8326	0.9430
	N=50	0.9436	0.7839	0.9458	0.9217	0.7352	0.9361	0.9065	0.7746	0.9259
$E[Y] \cong 1$		$\mu = 1$	$\nu = 0.8$	**E[Y] = 1.09**	$\mu = 0.8$	$\nu = 0.6$	**E[Y] = 1.07**	$\mu = 0.5$	$\nu = 0.4$	**E[Y] = 1.04**
	N=1000	0.9356	0.9035	0.8824	0.8698	0.7838	0.8556	0.6707	0.6170	0.8411
	N=100	0.4931	0.5664	0.7478	0.3513	0.4766	0.6618	0.1966	0.4012	0.6922
	N=50	0.1199	0.3260	0.6343	-0.1220	0.2147	0.5700	-0.2445	0.1184	0.5460
$E[Y] \cong 0.5$		$\mu = 0.5$	$\nu = 0.8$	**E[Y] = 0.53**	$\mu = 0.3$	$\nu = 0.6$	**E[Y] = 0.47**	$\mu = 0.14$	$\nu = 0.4$	**E[Y] = 0.46**
	N=1000	0.6891	0.5173	0.4843	0.5035	0.4607	0.5032	0.2539	0.3292	0.4754
	N=100	0.3944	0.4145	0.5344	0.2109	0.2854	0.4038	-0.0107	0.1639	0.4424
	N=50	-0.1546	0.0278	0.4100	-0.1432	-0.6343	0.3039	-0.2790	-0.1516	0.3098

From the results of Table 6.7, for the sample mean value of 10, it is clear that a sample size of less than 50 is sufficient for proper estimation of parameters if the threshold level of 85% is considered. When the sample mean value equals 1, a sample size of nearly 1,000 is sufficient for proper posterior estimation of parameters. For the sample mean value of 0.5, a sample size much greater than 1000 is needed. The exact sample size is known by performing extra simulation with different sample sizes.

6.2.2 Gamma prior

This section presents the results for the assumption of gamma prior for the shape parameter during the posterior estimation.

Table 6.8 presents the simulation results for the sample mean of 10. When these results are compared to the results of Table 6.2, there is no noticeable difference. The assumption of different priors for the shape parameter has not much influence on the posterior estimates for the sample mean of 10. Similar to the results in Table 6.2, the parameters are accurately estimated and are close to their theoretical value. As the sample decreased, the standard deviation of the estimates increased, meaning that the confidence interval in the prediction of the estimates becomes wider with the decrease in sample size.

Table 6.8: Results of parameters for $E[Y] \cong 10$ (gamma prior)

	N=1000			N=1000			N=1000		
	$\mu = 10$	**$\nu = 0.8$**	**E[Y]=10.05**	**$\mu = 10$**	**$\nu = 0.6$**	**E[Y]=10.14**	**$\mu = 10$**	**$\nu = 0.4$**	**E[Y]=10.55**
Mean	9.8878	0.7794	9.9043	9.7949	0.6164	9.9183	9.8185	0.4230	10.3199
Std.dev	0.0968	0.0459	0.0887	0.1115	0.0307	0.1110	0.1552	0.0221	0.1258
Min.	9.5196	0.6836	9.5978	9.4955	0.5434	9.6475	9.4393	0.3706	10.0058
Max.	10.1800	0.8902	10.2060	10.0446	0.6725	10.1642	10.2181	0.4702	10.5504
	N=100			N=100			N=100		
Mean	9.8103	0.8009	9.8334	9.7722	0.6128	9.9435	9.7460	0.4260	10.2756
Std.dev	0.3979	0.1162	0.3942	0.4212	0.0994	0.3726	0.5344	0.0661	0.4572
Min.	8.8335	0.5472	8.8131	8.9412	0.4024	9.1332	8.3141	0.2887	8.9438
Max.	10.7120	1.1466	10.7333	10.8118	0.9252	10.8637	11.1169	0.6417	11.6668
	N=50			N=50			N=50		
Mean	9.7814	0.7757	9.8405	9.6683	0.6069	9.8728	9.6551	0.4157	10.2560
Std.dev	0.5729	0.2000	0.5525	0.6360	0.1356	0.5756	0.8375	0.0917	0.7009
Min.	8.2246	0.3980	8.2011	8.2901	0.2980	8.6048	7.3619	0.1935	8.3725
Max.	10.8595	1.6108	10.9561	11.1157	1.1931	11.0261	11.8657	0.6317	12.0172

The simulation results of the parameter estimates for the sample mean of 1 are presented in Table 6.9. The results presented in the table are much similar to the results in Table 6.3. The parameters are mis-estimated for all sample sizes and the standard deviation became highly noticeable for the sample size of 50. The centering parameter for n=50 and $\nu = 0.4$ is not significantly different from zero.

Table 6.10 presents the simulation results for the sample mean of 0.5. The parameter estimated are highly unreliable and biased for the sample mean of 0.5, similar to the results presented in Table 6.4. In most the cases, the minimum value of the estimates is negligible and is almost equal to zero. Even for low sample mean value, the assumption of different priors for the shape parameter has a negligible effect.

Table 6.9: Results of parameters for $E[Y] \cong 1$ (gamma prior)

	N=1000			N=1000			N=1000		
	$\mu = 1$	$\nu = 0.8$	**E[Y]=1.09**	$\mu = 0.8$	$\nu = 0.6$	**E[Y]=1.07**	$\mu = 0.5$	$\nu = 0.4$	**E[Y]=1.04**
Mean	1.0046	0.8585	0.9742	0.9025	0.7654	0.9434	0.7012	0.6199	0.9079
Std.dev	0.0562	0.0511	0.0226	0.0643	0.0433	0.0300	0.1024	0.0763	0.0541
Min.	0.8206	0.7100	0.9135	0.7206	0.6430	0.8753	0.2951	0.3227	0.8252
Max.	1.1536	1.0115	1.0382	1.0371	0.8699	1.0743	0.9711	0.7790	1.0743
	N=100			N=100			N=100		
Mean	0.8529	0.7855	0.9723	0.6544	0.6209	0.9419	0.4041	0.4150	0.9219
Std.dev	0.2849	0.2569	0.1098	0.3282	0.2889	0.2025	0.2953	0.2480	0.2834
Min.	0.0387	0.1375	0.7346	0.0000	0.0670	0.0000	0.0000	0.0612	0.0000
Max.	1.4918	1.3930	1.2179	1.2969	1.4507	1.4971	1.0807	1.1991	1.5954
	N=50			N=50			N=50		
Mean	0.5987	0.6211	0.9276	0.5203	0.5148	0.9405	0.3355	0.3670	0.9118
Std.dev	0.4127	0.3927	0.2593	0.3793	0.3340	0.3185	0.3624	0.3151	0.3624
Min.	0.0000	0.0654	0.0000	0.0000	0.0609	0.0000	0.0000	0.0637	0.0000
Max.	1.3179	1.5501	1.5528	1.3733	1.4851	1.6097	1.3391	1.4114	1.7847

Table 6.10: Results of parameters for $E[Y] \cong 0.5$ (gamma prior)

	N=1000			N=1000			N=1000		
	$\mu = 0.5$	$\nu = 0.8$	**E[Y]=0.53**	$\mu = 0.3$	$\nu = 0.6$	**E[Y]=0.47**	$\mu = 0.14$	$\nu = 0.4$	**E[Y]=0.46**
Mean	0.7132	1.5243	0.3478	0.5880	1.2920	0.3158	0.5346	1.1958	0.3016
Std.dev	0.0428	0.0858	0.0124	0.0576	0.1029	0.0192	0.0540	0.0910	0.0126
Min.	0.5909	1.2788	0.3174	0.4150	1.0020	0.2789	0.3745	0.9395	0.2683
Max.	0.8083	1.7288	0.3789	0.7352	1.5789	0.4459	0.6847	1.4765	0.3375
	N=100			N=100			N=100		
Mean	0.4538	1.0153	0.4078	0.3064	0.7768	0.3711	0.2460	0.6810	0.3576
Std.dev	0.2554	0.5096	0.1269	0.2373	0.5014	0.1406	0.2108	0.5039	0.1426
Min.	0.0000	0.1101	0.0000	0.0000	0.1125	0.0000	0.0000	0.1118	0.0000
Max.	0.9744	2.1478	0.6481	0.8458	2.0120	0.7447	0.7448	3.3746	0.7564
	N=50			N=50			N=50		
Mean	0.3524	0.8751	0.4145	0.2444	0.6935	0.3604	0.2090	0.6213	0.3614
Std.dev	0.2772	0.6428	0.1632	0.2321	0.5683	0.1771	0.2215	0.5397	0.1843
Min.	0.0000	0.1010	0.0000	0.0000	0.0997	0.0000	0.0000	0.0981	0.0000
Max.	0.9125	2.4002	0.8291	0.9295	3.1123	0.7732	0.8567	2.3918	0.7740

Tables 6.11 and 6.12 present the bias and MSE of the parameter estimation respectively. Although a clear systematic bias is not seen with the decrease in sample size, in almost all the cases, the MSE is always high for the sample size equal to 50. In general, as the sample size and sample mean decreased, the MSE increased. The results presented here are much similar to the results presented in Tables 6.5 and 6.6.

Table 6.11: Bias in the parameter estimation (gamma prior)

$E[Y] \cong 10$		$\mu = 10$	$\nu = 0.8$	E[Y] = 10.05	$\mu = 10$	$\nu = 0.6$	E[Y] = 10.14	$\mu = 10$	$\nu = 0.4$	E[Y] = 10.55
	N=1000	0.1122	0.0206	0.1457	0.2051	-0.0164	0.2217	0.1815	-0.0230	0.2301
	N=100	0.1897	-0.0009	0.2166	0.2278	-0.0128	0.1965	0.2540	-0.0260	0.2744
	N=50	0.2186	0.0243	0.2095	0.3317	-0.0069	0.2672	0.3449	-0.0157	0.2940
$E[Y] \cong 1$		$\mu = 1$	$\nu = 0.8$	E[Y] = 1.09	$\mu = 0.8$	$\nu = 0.6$	E[Y] = 1.07	$\mu = 0.5$	$\nu = 0.4$	E[Y] = 1.04
	N=1000	-0.0046	-0.0585	0.1158	-0.1025	-0.1654	0.1266	-0.2012	-0.2199	0.1321
	N=100	0.1471	0.0145	0.1177	0.1456	-0.0209	0.1281	0.0959	-0.0150	0.1181
	N=50	0.4013	0.1789	0.1624	0.2797	0.0852	0.1295	0.1645	0.0330	0.1282
$E[Y] \cong 0.5$		$\mu = 0.5$	$\nu = 0.8$	E[Y] = 0.53	$\mu = 0.3$	$\nu = 0.6$	E[Y] = 0.47	$\mu =$ 0.14	$\nu = 0.4$	E[Y] = 0.46
	N=1000	-0.2132	-0.7243	0.1822	-0.2880	-0.6920	0.1542	-0.3946	-0.7958	0.1584
	N=100	0.0462	-0.2153	0.1222	-0.0064	-0.1768	0.0989	0.2240	-0.2110	0.1024
	N=50	0.1476	-0.3751	0.1155	0.2556	-0.1935	0.1096	0.2910	-0.1213	0.0986

Table 6.12: MSE of the estimated parameters (gamma prior)

$E[Y] \cong 10$		$\mu = 10$	$\nu = 0.8$	E[Y] = 10.05	$\mu = 10$	$\nu = 0.6$	E[Y] = 10.14	$\mu = 10$	$\nu = 0.4$	E[Y] = 10.55
	N=1000	0.0220	0.0025	0.0291	0.0545	0.0012	0.0615	0.0570	0.0010	0.0688
	N=100	0.1943	0.0135	0.2023	0.2293	0.0101	0.1774	0.3501	0.0050	0.2843
	N=50	0.3759	0.0406	0.3492	0.5145	0.0184	0.4027	0.8203	0.0087	0.5777
$E[Y] \cong 1$		$\mu = 1$	$\nu = 0.8$	E[Y] = 1.09	$\mu = 0.8$	$\nu = 0.6$	E[Y] = 1.07	$\mu = 0.5$	$\nu = 0.4$	E[Y] = 1.04
	N=1000	0.0032	0.0060	0.0139	0.0146	0.0292	0.0169	0.0509	0.0542	0.0204
	N=100	0.1028	0.0662	0.0259	0.1289	0.0839	0.0574	0.0964	0.0617	0.0942
	N=50	0.3314	0.1862	0.0936	0.2221	0.1188	0.1182	0.1584	0.1004	0.1478
$E[Y] \cong 0.5$		$\mu = 0.5$	$\nu = 0.8$	E[Y] = 0.53	$\mu = 0.3$	$\nu = 0.6$	E[Y] = 0.47	$\mu =$ 0.14	$\nu = 0.4$	E[Y] = 0.46
	N=1000	0.0473	0.5320	0.0333	0.0863	0.4894	0.0242	0.1586	0.6415	0.0253
	N=100	0.0673	0.3060	0.0310	0.0563	0.2827	0.0296	0.0946	0.2984	0.0308
	N=50	0.0986	0.5540	0.0400	0.1192	0.3604	0.0434	0.1337	0.3059	0.0437

As mentioned above, the cutoff factor is used for knowing recommended sample size for a given sample mean. Table 6.13 gives the cutoff values for the parameters with the assumption of gamma prior for the shape parameter. In this research, the cutoff value for the sample mean for $\nu = 0.4$ is used for determining the sample size. With the threshold value of 85%, the sample size less than 50 is sufficient for the sample mean of 10. The exact sample size for the sample mean of 10 will be determined by performing more simulation runs (not shown here). For the sample mean of 1, the sample size of 1000 is sufficient since the cutoff value is almost equal to the threshold value. For the sample mean of 0.5, a sample size much greater than 1000 is required to reduce unreliable estimation of parameters.

Table 6.13: Cutoff factor of the estimated parameters (gamma prior)

$E[Y] \cong 10$		**$\mu = 10$**	**$\nu = 0.8$**	**E[Y] = 10.05**	**$\mu = 10$**	**$\nu = 0.6$**	**E[Y] = 10.14**	**$\mu = 10$**	**$\nu = 0.4$**	**E[Y] = 10.55**
	N=1000	0.9850	0.9354	0.9828	0.9762	0.9436	0.9750	0.9757	0.9246	0.9746
	N=100	0.9551	0.8550	0.9543	0.9510	0.8364	0.9576	0.9393	0.8333	0.9481
	N=50	0.9373	0.7403	0.9400	0.9258	0.7762	0.9357	0.9062	0.7761	0.9259
$E[Y] \cong 1$		**$\mu = 1$**	**$\nu = 0.8$**	**E[Y] = 1.09**	**$\mu = 0.8$**	**$\nu = 0.6$**	**E[Y] = 1.07**	**$\mu = 0.5$**	**$\nu = 0.4$**	**E[Y] = 1.04**
	N=1000	0.9439	0.9095	0.8789	0.8659	0.7766	0.8620	0.6781	0.6245	0.8428
	N=100	0.6241	0.6724	0.8344	0.4514	0.5335	0.7456	0.2317	0.4013	0.6670
	N=50	0.0385	0.3053	0.6702	0.0941	0.3303	0.6345	-0.1865	0.1367	0.5784
$E[Y] \cong 0.5$		**$\mu = 0.5$**	**$\nu = 0.8$**	**E[Y] = 0.53**	**$\mu = 0.3$**	**$\nu = 0.6$**	**E[Y] = 0.47**	**$\mu = 0.14$**	**$\nu = 0.4$**	**E[Y] = 0.46**
	N=1000	0.6951	0.5215	0.4751	0.5005	0.4585	0.5079	0.2550	0.3302	0.4730
	N=100	0.4281	0.4551	0.5679	0.2254	0.3155	0.5368	-0.2500	0.1979	0.5091
	N=50	0.1086	0.1495	0.5178	-0.4130	0.1343	0.4222	-0.7496	0.1097	0.4218

Similar to the results with log-normal prior, Table 6.13 shows that a sample size less than 50 is sufficient for the sample mean value equal to 10. When sample mean equals 1, a sample size of 1,000 is sufficient for proper estimation of parameters. The sample size of much greater than 1,000 is needed when the data sample mean value equals 0.5.

6.2.3 Recommended sample size

Table 6.14 gives the recommended sample size for developing the COM-Poisson models. The recommended sample size are necessary to circumvent the problem of low sample mean and small sample size. If the sample size is less than the recommended sample size, then the parameter estimates is likely to become unreliable and biased. The sample sizes recommended below were confirmed by performing additional simulations (shown in Appendix A).

Table 6.14: Recommended sample size for minimizing the unreliable estimation of parameters

Sample mean $E[Y]$	Minimum Sample size
10	25
5	50
2	150
1	1000
0.75	3000
0.5	5000

6.3 Discussion

The results of this study show that the COM-Poisson models are affected by the low sample mean and small sample size. The results of the simulation study presented above raise few important points that merit further discussion.

- First, the parameters of COM-Poisson models are estimated accurately irrespective of the sample sizes for a sample mean of 10. Although the estimated values are close to the theoretical value, the standard deviation of the estimates became large for the sample size of 100 and 50. The difference between the minimum and maximum value of the estimates increased as the sample size decreased.
- Second, for the sample mean of 1 and 0.5, the parameter estimates started deviating from the theoretical value. When the mean is equal to 0.5 and sample size is 50, the estimates are highly unreliable and biased. Sometimes, the estimates are not significantly different from zero at these extreme cases. But it is clear from the simulation results that there is no systematic bias in the prediction of estimates.
- Third, the log-normal prior or a gamma prior on the shape parameter did not have different effect on the results of parameter estimates.

6.4 Summary

The objective of the study in this chapter was to examine the effects of low sample mean and small sample size on the parameter estimates of COM-Poisson models. The simulation study was used to quantify the prediction accuracy. Three different sample means: E[Y] $\cong$ 10, 1.0 and 0.5, and three different sample sizes: *n*= 1000, 100, 50 were considered in the analysis. During posterior analysis, the estimation was carried out with a log-normal prior and a gamma prior for the shape parameter.

The simulation results show that the parameter estimates are very close to the theoretical value for all the parameters of COM-Poison models for a sample mean of 10 irrespective of the sample sizes. However, the standard deviation increased with the decrease in the sample sizes. For the sample mean equal to 1 and 0.5, the parameters are mis-estimated. Although the posterior mean is close to the theoretical value for the sample mean of 1, the standard deviation and the difference between minimum and maximum value became larger compared to the sample mean equal to 10. For the sample mean of 0.5 and the sample size of 50 (i.e., extreme conditions), the estimates are highly unreliable and biased. Also, the assumption of either a log-normal prior or a gamma prior had a similar effect on the parameter estimates.

CHAPTER VII

SUMMARY AND CONCLUSIONS

There has been considerable research in the traffic safety literature which deals with the development of statistical models for analyzing motor vehicle crashes. The most common probabilistic structure of the models used by transportation safety analysts for modeling motor vehicle crashes are the traditional Poisson and Negative Binomial distributions. Although the Poisson and NB regression models possess desirable distributional properties to describe motor vehicle accidents, these models are not without limitations. These limitations include the biased goodness-of-fit statistics, improper estimation of dispersion parameter and biased parameter estimates when the crash data are characterized by low sample mean and small sample size. The other important limitation associated with NB models is the mis-specification of the probability density function (PDF) when the data exhibits under-dispersion, the condition that the mean is greater than the variance. Many new statistical methods have been proposed to overcome the difficulties that are raised by traditional Poisson and Poisson-gamma models. None of these new models were able to replace the NB models for analyzing traffic crash data.

The primary objectives of this research were, to characterize the performance of COM-Poisson distribution (an innovative distribution which is an extension of Poisson distribution), to introduce COM-Poisson GLMs for modeling different traffic crash datasets and finally to quantify the effect of small sample size and low sample mean on COM-Poisson models.

This chapter first presents the summary of the research work and then describes proposed directions for future research work.

7.1 Summary of Work

This section briefly presents the major contributions of this research. The contributions of this dissertation include characterizing the properties of the COM-Poisson distribution and knowing how this new distribution and its GLM could be used by the traffic safety community.

7.1.1 Performance of the COM-Poisson GLM

As discussed in Chapters II and III, though the COM-Poisson distribution was introduced a few decades ago, the statistical and probabilistic properties of this distribution were derived only recently. Although it was known that the COM-Poisson distribution handles count data, it is important to better understand its performance for the wide variety of situations. This need motivated to do a research study on this topic presented in Chapter III. For the purpose of this research, nine different scenarios were evaluated. These scenarios included under-, equi- and over-dispersed datasets with low, moderate and high sample mean values respectively. The simulation study was used to assess the performance of COM-Poisson distribution.

The true (assigned) parameters for the centering and shape parameters were compared with the posterior estimates of MCMC runs using with graphical plots. The results of this study indicated that the true parameters is located in the 95% credible interval for nearly all scenarios and are generally close to the estimated posterior mean of the parameters. Though the true parameters were inside the credible intervals for all scenarios, these intervals were found to be wider for the low mean values for both the centering and shape parameters. The bias in the prediction of the parameters and the mean value also increased as the data sample mean values decreased. Even for the low sample mean values, the bias was considerably less for under-dispersed datasets than for over-dispersed and equi-dispersed datasets. The other important finding from this study showed that despite its flexibility in handing count data with all dispersions, the COM-Poisson distribution suffers from important limitations for low mean over-dispersed data.

The other motivation towards this study was to better understand how well the approximation for the mean and variance suggested by Shmueli et al (2005) works. It was found that the asymptotic approximation of the mean approximates the true mean adequately even for $E[Y] > 5$. This value determined through numerical analysis of the COM-Poisson GLM was substantially lower than the lower bound value equal to 10 suggested previously. The accuracy of the approximation dropped as the sample mean value decreased. The asymptotic approximation was accurate for all datasets with high and moderate sample mean values irrespective of the dispersion in the data. The approximation was also accurate for low sample mean values for under-dispersed datasets. However, the accuracy dropped substantially for low sample mean values for over-dispersed and equi-dispersed datasets.

Finally, it was also found that the datasets with higher sample mean values required more computational time for a given number of replications than the low mean datasets did. Similarly, it is important to note that the over-dispersed datasets required more computational time than the other type of datasets.

7.1.2 Application to over-dispersed crash data

The second part of this dissertation documented in Chapter IV was related to the application of the COM-Poisson GLM for analyzing motor vehicle crashes. This study was motivated by the fact that many researchers shifted their interests in applying new methods for analyzing crash data because of limitations associated with the commonly used NB models. The application of COM-Poisson for analyzing crash data was first investigated and then compared with the NB models. The comparison analysis was carried out using the most common functional forms used by transportation safety analysts, which link crashes to the entering flows at intersections or on segments. This comparison was important since the foremost objective was to find whether COM-Poisson GLM could replace NB models for analyzing traffic crash data.

Using 4-legged signalized intersections crash data collected in Toronto and rural 4-lane divided and undivided highways crash data collected in Texas, several full Bayes (FB)

NB and COM-Poisson GLMs were developed. Several methods were used to assess the statistical fit and predictive performance of the models. The results of this study showed that COM-Poisson GLMs perform as well as FB NB models in terms of GOF statistics and predictive performance. The estimated values with COM-Poisson and NB models are slightly different, with the COM-Poisson output being always lower than the NB output.

The important point noted, as documented in the literature, is that the NB GLM could theoretically handle under-dispersion, since the dispersion parameter can be negative ($Var\left(Y\right)=\mu+(-\alpha)\mu^{2}$). However, in this case, the mean of the Poisson is no longer gamma distributed because this latter distribution cannot have negative parameters (i.e., $gamma\left(\phi,\phi\right)$). In addition, researchers who have worked on the characterization of the NB distribution and GLM have indicated that a negative dispersion parameter could lead to a mis-specification of the PDF (when $-1/\left(\text{max of counts}\right)<\alpha$).

The relative marginal effects of the covariates were also investigated. It was found that the relative marginal effect of the major flow for the COM-Poisson model depended on the major and minor entering flows. The relative marginal effect on the expected mean for a unit increase in major flow remained nearly constant for lower minor flow volumes (e.g., 100 veh/day) and the rate of curvature increased with the increase in minor flows. On the other hand, the relative marginal effect of the major flow for the NB model was independent of the minor flow. With a unit increase in major flow, the relative marginal effect on the expected mean value decreased. This decrease was smaller for higher major flows. Similar results were seen for both COM-Poisson and NB models for the marginal effect related to the minor flow.

7.1.3 Application to under-dispersed crash data

It was clear from the results of the Chapter IV that the COM-Poisson distribution performs as well as the NB distribution. The next objective was to verify if COM-Poisson distribution outperforms NB distribution in some way. The research was motivated by the fact that the COM-Poisson models can handle under-dispersed data easily (as detailed in

Chapters II, III and IV) whereas the NB models have difficulties in handling under-dispersed data (as detailed in Chapter IV). In this research, a comparison was made between COM-Poisson model and traditional Poisson and gamma probability models. For each of these models, several methods were used to assess the statistical fit and predictive performance.

Using crash data collected at 162 railway-highway crossings in South Korea between 1998 and 2002, several COM-Poisson models were first estimated. The modeling results were compared to those produced from the Poisson and gamma probability models documented in Oh et al. (2006) and was documented in the Chapter V. This study has shown that the COM-Poisson models provide better statistical performance than the Poisson and gamma probability models when under-dispersion is observed in the data. Furthermore, the gamma probability model works as a dual-state model whereas COM-Poisson does not assume a dual-state data generating process. Thus the COM-Poisson GLM offers a more defensible approach for modeling under-dispersed data.

An interesting finding in this research was that with the inclusion of all eight significant variables, the COM-Poisson GLM shows that the modeling output exhibits near equi-dispersion. This characteristic has also been documented in Miaou and Song (2005), who proposed complex multivariate hierarchical predictive models for analyzing crash data. They showed that by significantly improving the mean function, one could almost eliminate the over-dispersion. In this study, it was also found that when the two non-significant variables were removed, the COM-Poisson model clearly showed signs of over-dispersion. On the other hand, when the same six variables used in the Poisson model were utilized for the COM-Poisson model, the modeling output showed signs of under-dispersion. This unusual characteristic (i.e., changes from under-dispersion to over-dispersion or vice-versa by including or removing explanatory variables) was attributed to the very low sample mean value (0.33 crashes per year) of the dataset. Under this condition, this illustrates that the predicted mean values can significantly influence the variation found in the data.

7.1.4 Effects of small sample size and low sample mean

The effect of small sample size and low sample mean value on the parameter estimates of COM-Poisson models was then investigated. It was already clear from the literature review in Chapter II that the NB models suffer from important limitations when the data are characterized by low sample mean values and small sample size. A simulation study was used to better understand the effects on biases and parameter estimation of the COM-Poisson distribution at these extreme conditions. The results of the simulation study showed that the COM-Poisson models are affected by the low sample mean value and small sample size. Also, it was found that there is no effect of the different prior distribution for the shape parameter on the posterior estimation and biases of the parameters.

The following are findings of this simulation study with the sample mean of 0.5, 1 and 10 and the sample sizes of 50, 100 and 1000:

1. There was no change in the prediction of parameter estimates of COM-Poisson models for different sample sizes for the sample mean of 10. However, the standard deviation of the estimates increased with the decrease in sample size, although the predicted value was close to the theoretical value in all cases. The range between the minimum and maximum value of the estimates increased as the sample size decreased.
2. It was found that the parameter estimates started deviating from the theoretical value for the sample mean of 1 and 0.5. At the mean of 0.5 and sample size of 50, the estimates are highly unreliable and biased. Sometimes, the estimates are not significantly different from zero at these extreme cases. There was large difference between the minimum and maximum values for the lower sample means. This difference increased with the decrease in sample size.
3. The assumption of different priors for the shape parameter for the posterior estimation was also investigated. Two different priors: log-normal prior and gamma prior were examined. It was found that there was no major difference in

the prediction of the estimates and the standard deviations with the change in prior assumption.

7.2 Directions for Future Research

The following are the recommendations for future research:

- As discussed in Chapter III, due to the high computational time and lack of readily available software, the analysis was restricted to five simulation runs (or datasets) for each scenario. It is recommended to conduct the simulations with large number of replications (around 100) for each of the nine scenarios. In addition to these simulations, it is recommended to conduct the research with datasets having sample mean value around 0.3. The RHX crash data documented in Chapter V had a sample mean value of 0.33.
- It was clear from the results of Chapter III that the approximation proposed by Shmueli et al., (2005) worked well for all dispersions in the data with high and moderate mean. The approximation was considerably inaccurate for datasets with low mean and over-, and equi-dispersion. It is recommended to extend this research by conducting a simulation study on the datasets with sample mean between 0.8 and 5. It will be helpful in knowing exact cutoff of the sample mean value where the approximation starts to deviate by a large amount.
- The EB method is now used commonly in highway safety analyzes for refining the parameter estimates, countermeasure analysis and hotspot identification (identifying road sites with an unacceptable high accident risk). The research can be extended by developing an EB modeling framework for the COM-Poisson model.
- There are different varieties of models for identifying hazardous sites (also called as hotspot identification). It is recommended to conduct the research by developing the methods about how to use COM-Poisson GLMs for identifying hazardous sites.

- There have been discussions about the assumption related to the fixed dispersion for NB models (Hauer, 2001; Hydecker and Wu, 2003; Geedipally and Lord, 2008). Recently, researchers have started using a covariate-dependent dispersion parameter for NB models for analyzing crash data. The COM-Poisson GLMS developed in Chapter IV and V used a fixed shape parameter (independent of covariates). Further research should be done to examine the effects of a covariate-dependent shape parameter on COM-Poisson GLMs similar to that of NB GLMs for analyzing crash data.
- The computational time for the posterior estimation of parameters was significantly long because of characteristics of the data (e.g., low mean) and the inclusion of large number of covariates in the study documented in Chapter V. It is anticipated, however, that the computational time will be decreased significantly when the maximum likelihood estimation of the COM-Poisson distribution is used. Despite the computational time advantage, the MLE does not provide the full posterior distributions for the regression parameters nor does it allow expert knowledge to be incorporated through the use of informative priors. Sellers and Shmueli (2008) developed a likelihood formulation for COM-Poisson distribution by the time this dissertation was written. It is recommended to estimate the parameters with MLE and then compare it with the estimates in Chapters IV and V.
- It is possible that the dataset used in Chapter V could contain intermingled over- and under-dispersed counts given the changes in the nature of the variance function when variables are included or excluded. It is recommended that the COM-Poisson models with a dual link (covariate dependent shape parameter) are estimated for datasets exhibiting similar characteristics as the one used in Chapter V to see if different sites show different levels of dispersion.
- The simulation analysis carried out in Chapter VI had covariate independent centering and shape parameters. The research can be extended by performing the simulation study on covariate-dependent parameters. Crash datasets having low sample mean and small sample size should be analyzed to validate the findings of the simulation study in Chapter VI.

- The procedure for correcting the bias of the parameters caused by low sample mean and small sample size should be developed for COM-Poisson models, similar to the one developed for NB models (Park and Lord, 2008).
- As documented in Chapter IV, depending upon the specification of the parameters μ and ν, the COM-Poisson model can predict more zeros than the NB model for the same mean value. Nonetheless, both models should not be used as a direct substitute to zero-inflated models (when they are warranted). As stated in Shmueli et al., (2005), the COM-Poisson distribution's structure allows for a variety of generalizations such as zero-inflated data and dependence. It is recommended to develop zero-inflated COM similar to zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models (not for traffic safety applications).
- It was found that the risk of a mis-estimated posterior mean caused by LSM and SSS can be greatly minimized when an appropriate non-vague prior distribution is used for NB models (Lord and Miranda-Moreno, 2007). A similar analysis comparing the influence of informative and non-informative prior can be done for COM-Poisson models.

REFERENCES

Abdel-Aty, M., Addella, M.F., 2004. Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes: generalized estimating equations for correlated data. Transportation Research Record 1897, 106-115.

Abdelwahab, H.T., Abdel-Aty, M.A., 2002. Artificial neural networks and logit models for traffic safety analysis of toll plazas. Transportation Research Record 1784, 115-125.

Airoldi, E.M., Anderson, A.G., Fienberg, S.E., Skinner, K.K., 2006. Who wrote Ronald Reagan's radio addresses?. Bayesian Analysis 1 (2), 289-320.

ABS-CBN news online. (http://www.abs-cbnnews.com/storypage.aspx?StoryId=47892, accessed on March 22, 2008).

Agrawal, R., Lord, D., 2006. Effects of sample size on the goodness-of-fit statistic and confidence intervals of crash prediction models subjected to low sample mean values. Transportation Research Record 1950, 35-43.

Boatwright, P., Borle, S., Kadane. J. B., 2003. A model of the joint distribution of purchase quantity and timing. Journal of the American Statistical Association 98, 564-572.

Borle, S., Boatwright, P., Kadane.J.B., 2006. The timing of bid placement and extent of multiple bidding: an empirical investigation using eBay online auctions. Statistical Science 21(2), 194-205.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK.

Carlin, B., Louis, T., 2000. Bayes and Empirical Bayes Methods for Data Analysis, second ed. Chapman and Hall, Newyork.

Casella, G., Berger, R., 2001. Statistical Inference, second ed. Duxbury, Pacific Grove, CA.

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

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

De Lapparent, M., 2005. Individual cyclists' probability distributions of severe/fatal crashes in large French urban areas. Accident Analysis & Prevention 37(6), 1086-1092.

Dean, C.B., 1994. Modified pseudo-likelihood estimator of the overdispersion parameter in Poisson mixture models. Journal of Applied Statistics 21(6), 523-532.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis & Prevention 40 (4), 1257-1266.

Devroye Luc, 1986. Non-uniform Random Variate Generation. Springer-Verlag, New York.

El-Basyouny, K., Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models. Presented at the 88th Annual Meeting to the Transportation Research Board, Washington DC.

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

Engineering Statistics Handbook. (http://www.itl.nist.gov/div898/handbook/apr/section4/apr412.htm, accessed on August 31, 2008).

Fitzpatrick, K., Lord, D., Park, B.-J., 2008. Horizontal curve accident modification factors with consideration of driveway density on rural, four-lane highways in Texas. Transportation Research Record, in press.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., Thomsen, L.K., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accident Analysis and Prevention 27 (1), 1–20.

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, in press.

Geedipally S., Guikema, S.D., Dhavala, S., Lord, D., 2008. Characterizing the performance of a Bayesian Conway-Maxwell Poisson GLM. Presented at the 2008 Joint Statistical Meetings, Denver, CO, August 3-7, 2008.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, second ed. Chapman and Hall/CRC, New York.

Guikema, S.D., Coffelt, J.P., 2008. A flexible count data regression model for risk analysis. Risk Analysis, in press.

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 & Prevention, 33(6), 799-808.

Hauer, E, Persaud, B.N., 1984. Problem of identifying hazardous locations using accident data. Transportation Research Record 975, 36-43.

Hauer, E., Persaud, B.N., 1987. How to estimate the safety of rail-highway grade crossings and the safety effects of warning devices. Transportation Research Record 1114, 131-140.

Heydecker, B.G., Wu, J., 2001. Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. Advances in Engineering Software 32, 859-869.

Hughes, W., Eccles, K., Harwood, D., Potts, I., Hauer, E., 2005. Development of a Highway Safety Manual. Appendix C: Highway Safety Manual Prototype Chapter: Two-Lane Highways. NCHRP Web Document 62 (Project 17-18(4)). Washington, D.C. (http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp w62.pdf, accessed October 2007).

Kadane, J.B., Shmueli, G., Minka, T.P., Borle, S., Boatwright, P., 2006. Conjugate analysis of the Conway–Maxwell–Poisson distribution. Bayesian Analysis 1, 363–374.

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

Lloyd-Smith, J.O., 2007. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS ONE 2 (2), e180. (http://www.pubmedcentral.nih.gov/ articlerender.fcgi? artid =1791715, accessed July 2007).

Lord, D., 2000. The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models. Ph.D. dissertation, Department of Civil Engineering, University of Toronto, Toronto, Ontario.

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

Lord, D., Bonneson, J.A., 2005. Calibration of predictive models for estimating the safety of ramp design configurations. Transportation Research Record 1908, 88–95.

Lord, D., Bonneson, J.A., 2007. Development of accident modification factors for rural frontage road segments in Texas. Transportation Research Record 2023, 20-27.

Lord, D., Manar, A., Vizioli, A., 2005a. Modeling crash-flow-density and crash-flow-v/c ratio for rural and urban freeway segments. Accident Analysis & Prevention 37 (1), 185–199.

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

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

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

Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., van Schalkwyk, I.,Ivan, J.N., Lyon, C., Jonsson, T., 2008a. Methodology for estimating the safety performance of multilane rural highways. NCHRP Web-Only Document 126, National Cooperation Highway Research Program, Washington, D.C. (http://onlinepubs.trb.org/onlinepubs/ nchrp/nchrp w126.pdf, accessed on June 3 2008).

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

Lord, D., Geedipally, S.R., Guikema, S., 2008c. Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting under-dispersion. Submitted for publication.

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

Maycock, G., Hall, R.D., 1984. Accidents at 4-arm roundabouts. TRRL Laboratory Report 1120. Transportation and Road Research Laboratory, Crowthorne, Berkshire, England.

Miaou, S-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Presented at the 73rd Annual Meeting of Transportation Research Board, Washington DC.

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

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

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

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., Lord, D., Fu, L., 2007. Evaluation of alternative hyper-priors for Bayesian road safety analysis. Paper 08-1788. In: Proceedings of the 84th Annual Meeting of the 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 & Prevention 39(3), 459-468.

Natarajan, R., McCulloch, C., 1998. Gibbs sampling with diffuse proper priors: a valid approach to data driven inference?. Journal of Computational and Graphical Statistics 7 (3) 267–277.

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 & 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, in press.

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

Payne, R.W. (Ed.), 2000. The Guide to Genstat. Lawes Agricultural Trust, Rothamsted Experimental Station, Oxford, UK.

Peltola H., 2000. Background and principles of the Finnish safety evaluation tool. TARVA 13th ICTCT workshop, Corfu, Greece.

Persaud, B., Lyon, C., Nguyen, T., 1999. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transportation Research Record 1665, 7-12.

Persaud B.N., Retting, R., Garder, P., Lord, D., 2001. Observational before-after study of U.S. roundabout conversions using the empirical Bayes method. Transportation Research Record 1751, 1–8.

Piegorsch, W.W., 1990. Maximum likelihood estimation for the negative binomial dispersion parameter. Biometrics 46, 863–867.

Poch, M., Mannering, F.L., 1996. Negative binomial analysis of intersection-accident frequency. Journal of Transportation Engineering 122 (2), 105–113.

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.

Rao, J., 2003. Small Area Estimation. John Wiley and Sons, West Sussex, England.

Ridout, M.S., Besbeas, P., 2004. An empirical model for underdispersed count data. Statistical Modelling 4, 77–89.

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

Salmon, P.M., Regan, M.A., Johnston, I., 2005. Human error and road transport: Phase one- A framework for an error tolerant road transport system. Australian Transport Safety Bureau, Victoria, Australia. (http://www.monash.edu.au/muarc/reports/muarc256.pdf , accessed November 2008)

SAS Institute Inc., 2002. Version 9 of the SAS System for Windows, SAS Institute, Cary, NC.

Shankar, V., Milton, J., Mannering, F.L., 1997. Modeling accident frequency as zero-altered probability processes: an empirical inquiry. Accident Analysis & Prevention 29 (6), 829–837.

Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergal, 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. (Available from http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml).

Smyth, G. K., 1989. Generalized linear models with varying dispersion. J.R. Statist. Soc. B 51, 47-60.

Telang, R., Boatwright, P. Mukhopadhyay, T., 2004. A mixture model for internet search-engine visits. Journal of Marketing 41, 206-214.

Toft, N., Innocent, G.T., Mellor, D.J., Reid, S.W.J., 2006. The gamma-Poisson model as a statistical method to determine if microorganisms are randomly distributed in a food matrix. Food Microbiology 23 (1), 90-94.

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

Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics 31 (2&3), 221-229.

Venables, W.N., Smith, D.M., The R Development Team, 2005. An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics. Insightful Corporation, Seattle, WA.

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, 275-289.

Washington, S.P., Karlaftis, M., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall, Boca Raton, FL.

Wedagama, D.M., Bird, R.N., Metcalf, A.V., 2006. The influence of urban landuse on non-motorised transport casualties. Accident Analysis & Prevention 38 (6), 1049-1057.

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

World Health Organization, 2004. World Report on Road Traffic Injury Prevention: Summary. World Health Organization, Geneva.

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

Ye, Z., Zhang, Y., Lord, D., 2008. Investigating goodness-of-fit statistics for generalized linear crash models with low sample mean values. Submitted for publication.

APPENDIX A

SIMULATION RESULTS FOR DETERMINING RECOMMENDED SAMPLE SIZE FOR MINIMIZING THE UNRELIABLE ESTIMATION OF PARAMETERS

Table A.1: Results of parameters for different sample mean values

$E[Y] \cong 10$ **($\mu = 10$, $\nu = 0.4$, E[Y]=10.55)**									
	N=20			N=25			N=30		
Mean	8.8182	0.4293	9.8369	8.9385	0.4211	9.9554	9.4660	0.4367	10.0537
Std.dev	2.5119	0.2225	1.6251	1.1755	0.1818	1.1468	1.0552	0.1311	0.9272
Min.	0.0000	0.0092	0.0000	4.3982	0.1631	4.9532	6.5344	0.2225	7.6459
Max.	13.0485	1.1868	13.0655	12.4199	0.9894	12.8348	11.8769	0.8593	12.4953
$E[Y] \cong 5$ **($\mu = 4.3$, $\nu = 0.4$, E[Y]=5.04)**									
	N=50			N=70			N=100		
Mean	3.9053	0.4255	4.7021	4.0902	0.4185	4.7842	4.1238	0.4156	4.8033
Std.dev	0.9552	0.1481	0.4937	0.6757	0.0930	0.4374	0.5321	0.0809	0.3558
Min.	0.7463	0.0907	3.3813	1.3704	0.1296	3.6987	2.6513	0.2131	4.0010
Max.	5.7709	0.8055	6.2277	5.5811	0.6345	6.0310	5.3873	0.6624	5.7769
$E[Y] \cong 2$ **($\mu = 1.3$, $\nu = 0.4$, E[Y]=2.001)**									
	N=100			N=150			N=200		
Mean	1.0353	0.3833	1.8570	1.1222	0.4029	1.8524	1.2146	0.4295	1.8484
Std.dev	0.4866	0.1613	0.2935	0.3817	0.1486	0.1894	0.3006	0.0986	0.1250
Min.	0.0000	0.0414	0.0000	0.4015	0.1412	1.3745	0.3879	0.1978	1.4910
Max.	1.9551	0.8080	3.0475	1.9695	0.7821	2.3524	1.9397	0.6840	2.1444
$E[Y] \cong 0.75$ **($\mu = 0.3$, $\nu = 0.4$, E[Y]=0.751)**									
	N=1500			N=1800			N=3000		
Mean	0.6172	0.8225	0.5998	0.6149	0.8201	0.5945	0.5778	1.0046	0.8631
Std.dev	0.0556	0.0575	0.0293	0.0493	0.0481	0.0186	0.0510	0.1975	0.0800
Min.	0.4467	0.5462	0.5596	0.4818	0.6363	0.5585	0.4544	0.7559	0.7098
Max.	0.7174	0.9201	0.7345	0.7406	0.9395	0.6938	0.7486	1.2981	1.0546
$E[Y] \cong 0.5$ **($\mu = 0.14$, $\nu = 0.4$, E[Y]=0.46)**									
	N=2000			N=2500			N=5000		
Mean	0.5421	1.1954	0.3043	0.5446	1.1962	0.3051	0.5355	1.1757	0.5515
Std.dev	0.0376	0.0628	0.0086	0.0346	0.0577	0.0080	0.0204	0.0336	0.0163
Min.	0.4550	1.0549	0.2854	0.4544	1.0489	0.2864	0.4926	1.1064	0.5489
Max.	0.6241	1.3366	0.3239	0.6162	1.3181	0.3215	0.5811	1.2519	0.7073

Table A.2: Bias in the parameter estimation

$E[Y] \cong 10$		**μ = 10**	**ν = 0.4**	**E[Y] = 10.55**
	N=20	1.1818	-0.0293	0.7031
	N=25	1.0615	-0.0211	0.5846
	N=30	0.5340	-0.0367	0.4863
$E[Y] \cong 5$		**μ = 4.3**	**ν = 0.4**	**E[Y] = 5.04**
	N=50	0.3947	-0.0255	0.3379
	N=70	0.2098	-0.0185	0.2558
	N=100	0.1762	-0.0156	0.2367
$E[Y] \cong 2$		**μ = 1.3**	**ν = 0.4**	**E[Y] = 2.001**
	N=100	0.2647	0.0167	0.1440
	N=150	0.1778	-0.0029	0.1486
	N=200	0.2647	0.0167	0.1440
$E[Y] \cong 0.75$		**μ = 0.3**	**ν = 0.4**	**E[Y] = 0.751**
	N=1500	-0.3172	-0.4225	0.1512
	N=1800	-0.3149	-0.4201	0.1565
	N=3000	-0.2778	-0.6046	-0.1121
$E[Y] \cong 0.5$		**μ = 0.14**	**ν = 0.4**	**E[Y] = 0.46**
	N=2000	-0.4021	-0.7954	0.1579
	N=2500	-0.4046	-0.7962	0.1571
	N=5000	-0.3955	-0.7757	-0.0893

Table A.3: MSE of the estimated parameters

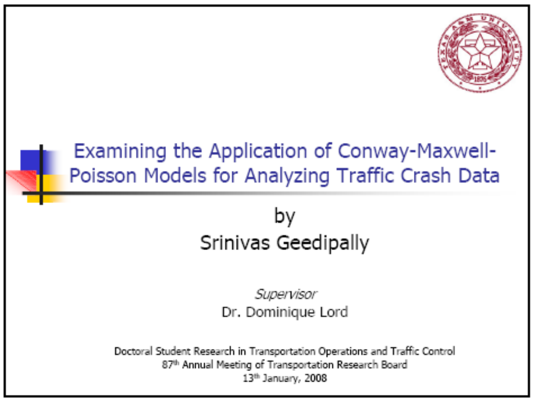
$E[Y] \cong 10$		**μ = 10**	**ν = 0.4**	**E[Y] = 10.55**
	N=20	7.7064	0.0504	3.1353
	N=25	2.5087	0.0335	1.6569
	N=30	1.3985	0.0185	1.0962
$E[Y] \cong 5$		**μ = 4.3**	**ν = 0.4**	**E[Y] = 5.04**
	N=50	1.0683	0.0226	0.3580
	N=70	0.5006	0.0090	0.2568
	N=100	0.3142	0.0068	0.1827
$E[Y] \cong 2$		**μ = 1.3**	**ν = 0.4**	**E[Y] = 2.001**
	N=100	0.3069	0.0263	0.1069
	N=150	0.1773	0.0221	0.0580
	N=200	0.0976	0.0106	0.0389
$E[Y] \cong 0.75$		**μ = 0.3**	**ν = 0.4**	**E[Y] = 0.751**
	N=1500	0.1037	0.1818	0.0237
	N=1800	0.1016	0.1788	0.0248
	N=3000	0.0798	0.4045	0.0190
$E[Y] \cong 0.5$		**μ = 0.14**	**ν = 0.4**	**E[Y] = 0.46**
	N=2000	0.1631	0.6365	0.0250
	N=2500	0.1649	0.6373	0.0247
	N=5000	0.1568	0.6029	0.0082

Table A.4: Cutoff factor of the estimated parameters

		$\mu = 10$	**$\nu = 0.4$**	**E[Y] = 10.55**
$E[Y] \cong 10$	N=20	0.6852	0.4773	0.8200
	N=25	0.8228	0.5653	***0.8707***
	N=30	0.8751	0.6882	0.8959
$E[Y] \cong 5$		**$\mu = 4.3$**	**$\nu = 0.4$**	**E[Y] = 5.04**
	N=50	0.7353	0.6468	***0.8728***
	N=70	0.8270	0.7734	0.8941
	N=100	0.8641	0.8018	0.9110
$E[Y] \cong 2$		**$\mu = 1.3$**	**$\nu = 0.4$**	**E[Y] = 2.001**
	N=100	0.4649	0.5768	0.8240
	N=150	0.6248	0.6313	***0.8700***
	N=200	0.7427	0.7604	0.8933
$E[Y] \cong 0.75$		**$\mu = 0.3$**	**$\nu = 0.4$**	**E[Y] = 0.751**
	N=1500	0.4783	0.4816	0.7432
	N=1800	0.4817	0.4844	0.7349
	N=3000	0.5112	0.3669	***0.8405***
$E[Y] \cong 0.5$		**$\mu = 0.14$**	**$\nu = 0.4$**	**E[Y] = 0.46**
	N=2000	0.2550	0.3326	0.4801
	N=2500	0.2544	0.3326	0.4845
	N=5000	0.2605	0.3396	***0.8354***

APPENDIX B

PRESENTATION AT "DOCTORAL STUDENT RESEARCH IN TRANSPORTATION OPERATIONS AND TRAFFIC CONTROL", 87TH ANNUAL MEETING OF TRANSPORTATION RESEARCH BOARD

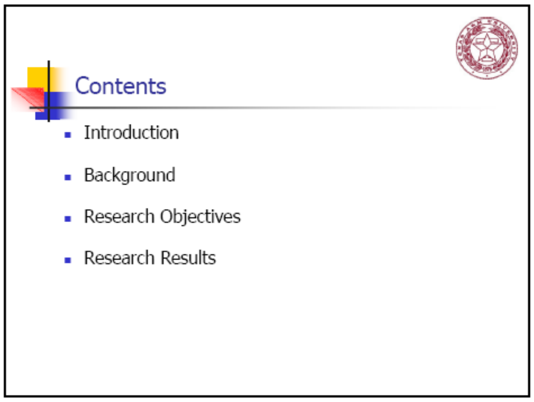


Examining the Application of Conway-Maxwell-Poisson Models for Analyzing Traffic Crash Data

by
Srinivas Geedipally

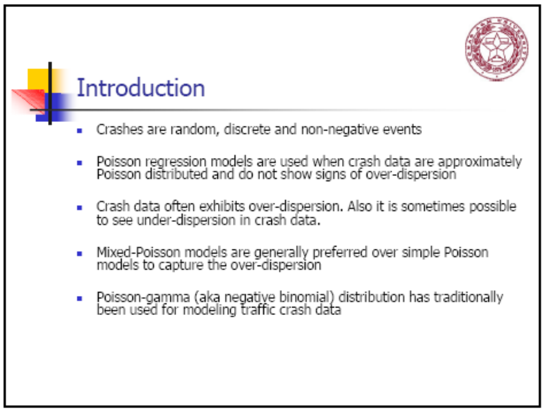
Supervisor
Dr. Dominique Lord

Doctoral Student Research in Transportation Operations and Traffic Control
87th Annual Meeting of Transportation Research Board
13th January, 2008



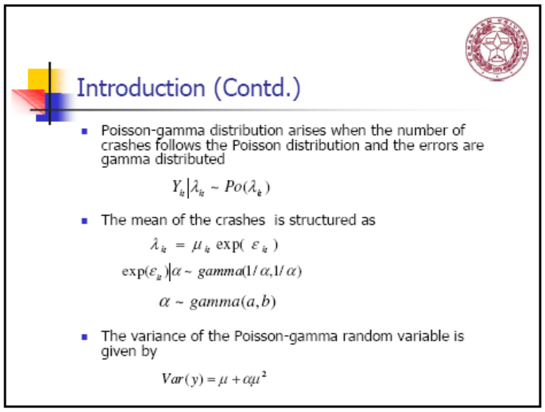
Contents

- Introduction
- Background
- Research Objectives
- Research Results



Introduction

- Crashes are random, discrete and non-negative events
- Poisson regression models are used when crash data are approximately Poisson distributed and do not show signs of over-dispersion
- Crash data often exhibits over-dispersion. Also it is sometimes possible to see under-dispersion in crash data.
- Mixed-Poisson models are generally preferred over simple Poisson models to capture the over-dispersion
- Poisson-gamma (aka negative binomial) distribution has traditionally been used for modeling traffic crash data



Introduction (Contd.)

- Poisson-gamma distribution arises when the number of crashes follows the Poisson distribution and the errors are gamma distributed

$$Y_{it}|\lambda_{it} \sim Po(\lambda_{it})$$

- The mean of the crashes is structured as

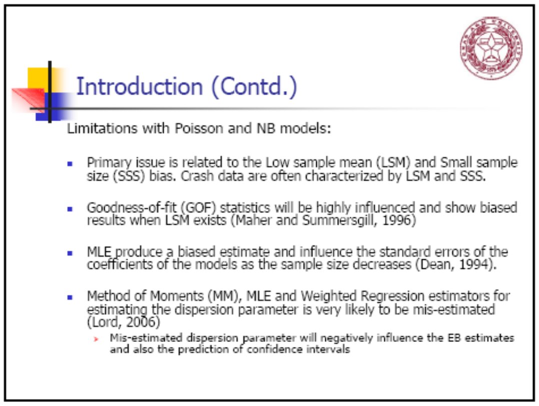
$$\lambda_{it} = \mu_{it}\exp(\varepsilon_{it})$$

$$\exp(\varepsilon_{it})|\alpha \sim gamma(1/\alpha, 1/\alpha)$$

$$\alpha \sim gamma(a,b)$$

- The variance of the Poisson-gamma random variable is given by

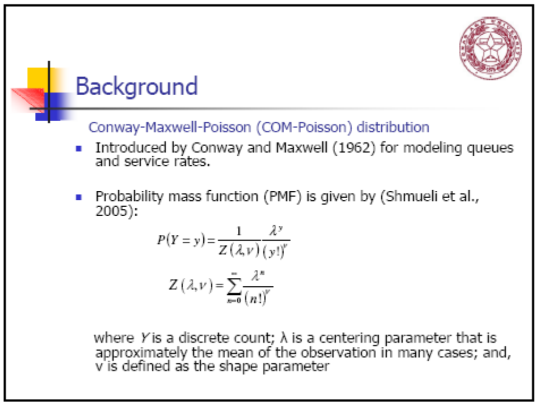
$$Var(y) = \mu + \alpha\mu^2$$



Introduction (Contd.)

Limitations with Poisson and NB models:

- Primary issue is related to the Low sample mean (LSM) and Small sample size (SSS) bias. Crash data are often characterized by LSM and SSS.
- Goodness-of-fit (GOF) statistics will be highly influenced and show biased results when LSM exists (Maher and Summersgill, 1996)
- MLE produce a biased estimate and influence the standard errors of the coefficients of the models as the sample size decreases (Dean, 1994).
- Method of Moments (MM), MLE and Weighted Regression estimators for estimating the dispersion parameter is very likely to be mis-estimated (Lord, 2006)
  - Mis-estimated dispersion parameter will negatively influence the EB estimates and also the prediction of confidence intervals



Background

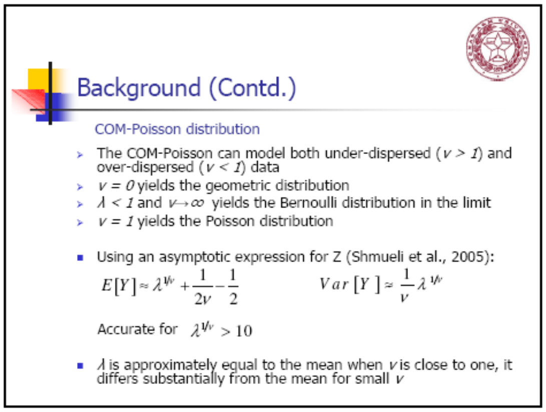
Conway-Maxwell-Poisson (COM-Poisson) distribution

- Introduced by Conway and Maxwell (1962) for modeling queues and service rates.
- Probability mass function (PMF) is given by (Shmueli et al., 2005):

$$P(Y=y)=\frac{1}{Z(\lambda,\nu)}\frac{\lambda^{y}}{(y!)^{\nu}}$$

$$Z(\lambda,\nu)=\sum_{n=0}^{\infty}\frac{\lambda^{n}}{(n!)^{\nu}}$$

where Y is a discrete count; λ is a centering parameter that is approximately the mean of the observation in many cases; and, v is defined as the shape parameter



Background (Contd.)

COM-Poisson distribution

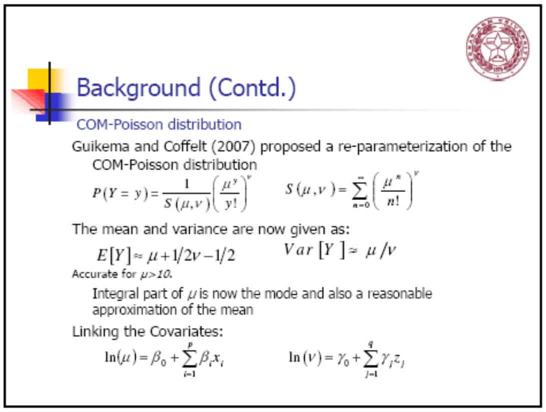
- The COM-Poisson can model both under-dispersed ($\nu > 1$) and over-dispersed ($\nu < 1$) data
- $\nu = 0$ yields the geometric distribution
- $\lambda < 1$ and $\nu \rightarrow \infty$ yields the Bernoulli distribution in the limit
- $\nu = 1$ yields the Poisson distribution

- Using an asymptotic expression for Z (Shmueli et al., 2005):

$$E[Y]\approx\lambda^{1/\nu}+\frac{1}{2\nu}-\frac{1}{2} \qquad Var[Y]\approx\frac{1}{\nu}\lambda^{1/\nu}$$

Accurate for $\lambda^{1/\nu} > 10$

- λ is approximately equal to the mean when ν is close to one, it differs substantially from the mean for small ν



Background (Contd.)

COM-Poisson distribution

Guikema and Coffelt (2007) proposed a re-parameterization of the COM-Poisson distribution

$$P(Y=y)=\frac{1}{S(\mu,\nu)}\left(\frac{\mu^{y}}{y!}\right)^{\nu} \qquad S(\mu,\nu)=\sum_{n=0}^{\infty}\left(\frac{\mu^{n}}{n!}\right)^{\nu}$$

The mean and variance are now given as:

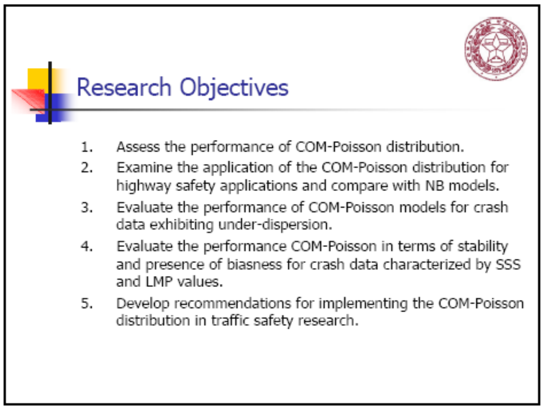
$$E[Y]\approx\mu+1/2\nu-1/2 \qquad Var[Y]\approx\mu/\nu$$

Accurate for $\mu > 10$.

Integral part of μ is now the mode and also a reasonable approximation of the mean

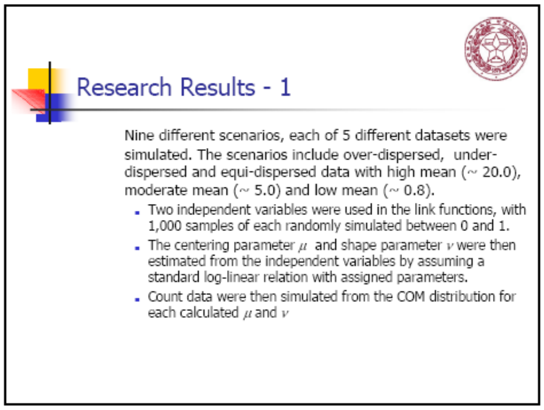
Linking the Covariates:

$$\ln(\mu)=\beta_0+\sum_{i=1}^{p}\beta_i x_i \qquad \ln(\nu)=\gamma_0+\sum_{j=1}^{q}\gamma_j z_j$$



Research Objectives

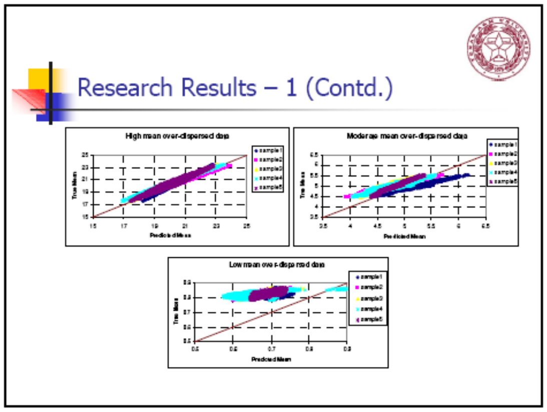
1. Assess the performance of COM-Poisson distribution.
2. Examine the application of the COM-Poisson distribution for highway safety applications and compare with NB models.
3. Evaluate the performance of COM-Poisson models for crash data exhibiting under-dispersion.
4. Evaluate the performance COM-Poisson in terms of stability and presence of biasness for crash data characterized by SSS and LMP values.
5. Develop recommendations for implementing the COM-Poisson distribution in traffic safety research.



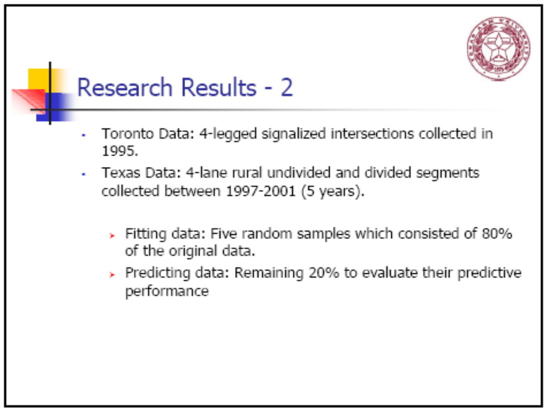
Research Results - 1

Nine different scenarios, each of 5 different datasets were simulated. The scenarios include over-dispersed, under-dispersed and equi-dispersed data with high mean (~ 20.0), moderate mean (~ 5.0) and low mean (~ 0.8).

- Two independent variables were used in the link functions, with 1,000 samples of each randomly simulated between 0 and 1.
- The centering parameter μ and shape parameter ν were then estimated from the independent variables by assuming a standard log-linear relation with assigned parameters.
- Count data were then simulated from the COM distribution for each calculated μ and ν

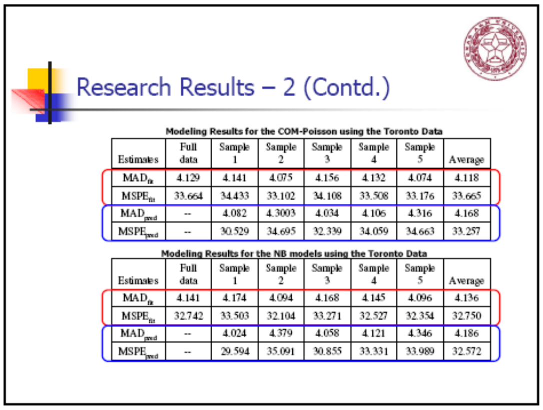


Research Results – 1 (Contd.)



Research Results - 2

- Toronto Data: 4-legged signalized intersections collected in 1995.
- Texas Data: 4-lane rural undivided and divided segments collected between 1997-2001 (5 years).
  - Fitting data: Five random samples which consisted of 80% of the original data.
  - Predicting data: Remaining 20% to evaluate their predictive performance



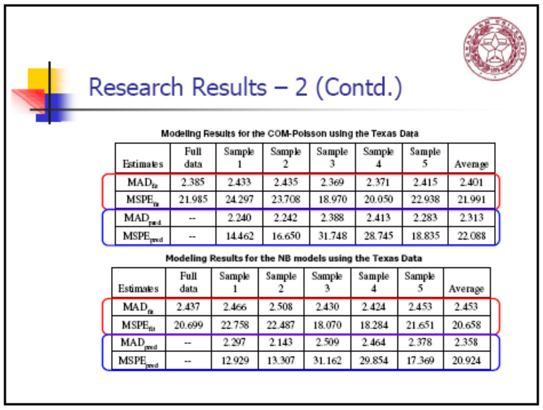
Research Results – 2 (Contd.)

Modeling Results for the COM-Poisson using the Toronto Data

Estimates	Full data	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average
MAD_{fit}	4.129	4.141	4.075	4.156	4.132	4.074	4.118
$MSPE_{fit}$	33.664	34.433	33.102	34.108	33.508	33.176	33.665
MAD_{pred}	--	4.082	4.3003	4.034	4.106	4.316	4.168
$MSPE_{pred}$	--	30.529	34.695	32.339	34.059	34.663	33.257

Modeling Results for the NB models using the Toronto Data

Estimates	Full data	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average
MAD_{fit}	4.141	4.174	4.094	4.168	4.145	4.096	4.136
$MSPE_{fit}$	32.742	33.503	32.104	33.271	32.527	32.354	32.750
MAD_{pred}	--	4.024	4.379	4.058	4.121	4.346	4.186
$MSPE_{pred}$	--	29.594	35.091	30.855	33.331	33.989	32.572



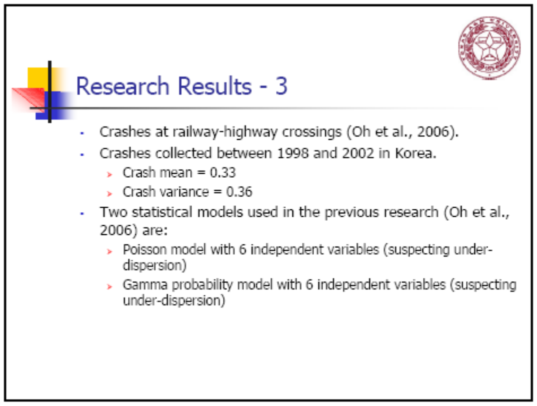
Research Results – 2 (Contd.)

Modeling Results for the COM-Poisson using the Texas Data

Estimates	Full data	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average
MAD_{fit}	2.385	2.433	2.435	2.369	2.371	2.415	2.401
$MSPE_{fit}$	21.985	24.297	23.708	18.970	20.050	22.938	21.991
MAD_{pred}	--	2.240	2.242	2.388	2.413	2.283	2.313
$MSPE_{pred}$	--	14.462	16.650	31.748	28.745	18.835	22.088

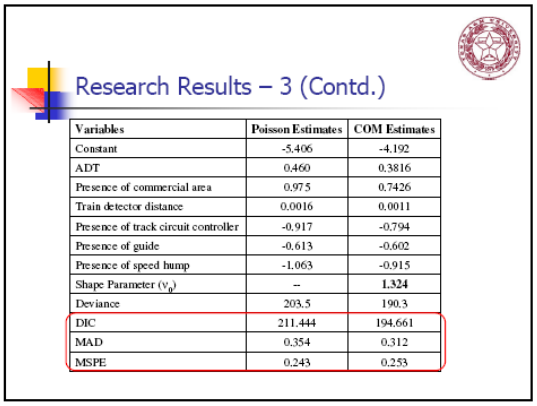
Modeling Results for the NB models using the Texas Data

Estimates	Full data	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average
MAD_{fit}	2.437	2.466	2.508	2.430	2.424	2.453	2.453
$MSPE_{fit}$	20.699	22.758	22.487	18.070	18.284	21.651	20.658
MAD_{pred}	--	2.297	2.143	2.509	2.464	2.378	2.358
$MSPE_{pred}$	--	12.929	13.307	31.162	29.854	17.369	20.924



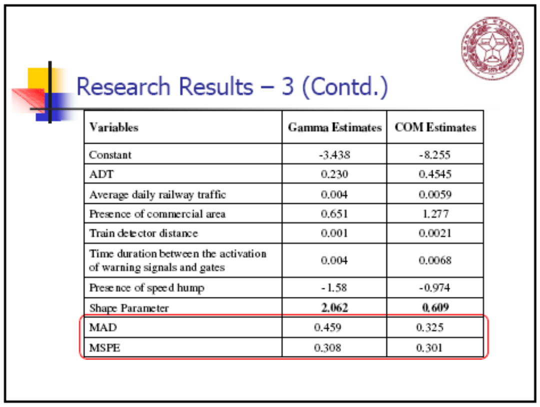
Research Results - 3

- Crashes at railway-highway crossings (Oh et al., 2006).
- Crashes collected between 1998 and 2002 in Korea.
  - Crash mean = 0.33
  - Crash variance = 0.36
- Two statistical models used in the previous research (Oh et al., 2006) are:
  - Poisson model with 6 independent variables (suspecting under-dispersion)
  - Gamma probability model with 6 independent variables (suspecting under-dispersion)



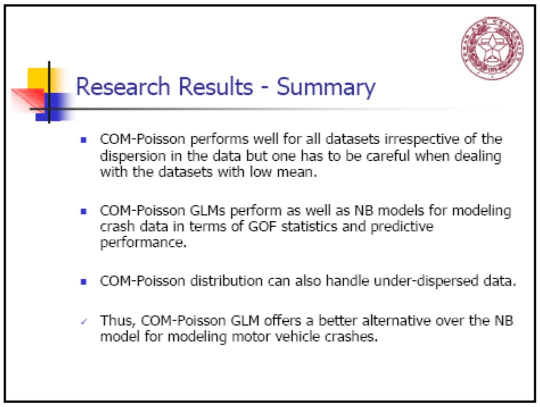
Research Results – 3 (Contd.)

Variables	Poisson Estimates	COM Estimates
Constant	-5.406	-4.192
ADT	0.460	0.3816
Presence of commercial area	0.975	0.7426
Train detector distance	0.0016	0.0011
Presence of track circuit controller	-0.917	-0.794
Presence of guide	-0.613	-0.602
Presence of speed hump	-1.063	-0.915
Shape Parameter (v_0)	--	**1.324**
Deviance	203.5	190.3
DIC	211.444	194.661
MAD	0.354	0.312
MSPE	0.243	0.253



Research Results – 3 (Contd.)

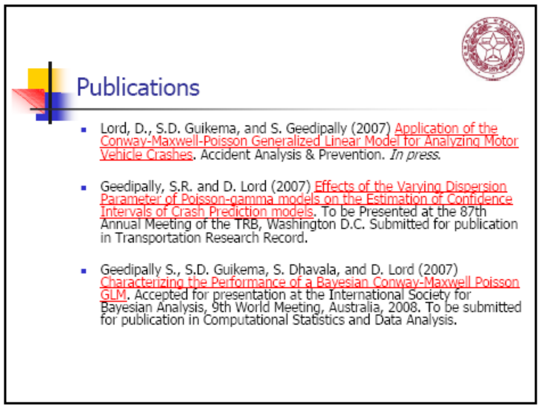
Variables	Gamma Estimates	COM Estimates
Constant	-3.438	-8.255
ADT	0.230	0.4545
Average daily railway traffic	0.004	0.0059
Presence of commercial area	0.651	1.277
Train detector distance	0.001	0.0021
Time duration between the activation of warning signals and gates	0.004	0.0068
Presence of speed hump	-1.58	-0.974
Shape Parameter	**2.062**	**0.609**
MAD	0.459	0.325
MSPE	0.308	0.301



Research Results - Summary

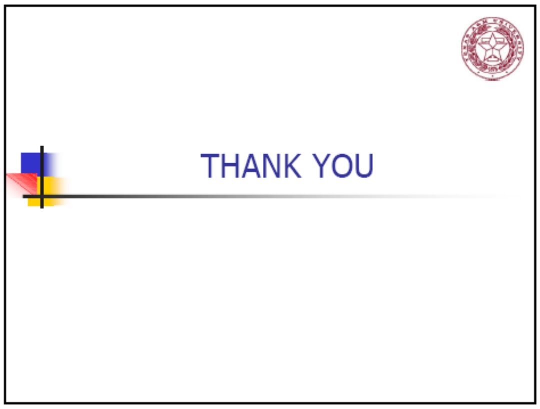
- COM-Poisson performs well for all datasets irrespective of the dispersion in the data but one has to be careful when dealing with the datasets with low mean.
- COM-Poisson GLMs perform as well as NB models for modeling crash data in terms of GOF statistics and predictive performance.
- COM-Poisson distribution can also handle under-dispersed data.

✓ Thus, COM-Poisson GLM offers a better alternative over the NB model for modeling motor vehicle crashes.



Publications

- Lord, D., S.D. Guikema, and S. Geedipally (2007) Application of the Conway-Maxwell-Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes. Accident Analysis & Prevention. *In press.*
- Geedipally, S.R. and D. Lord (2007) Effects of the Varying Dispersion Parameter of Poisson-gamma models on the Estimation of Confidence Intervals of Crash Prediction models. To be Presented at the 87th Annual Meeting of the TRB, Washington D.C. Submitted for publication in Transportation Research Record.
- Geedipally S., S.D. Guikema, S. Dhavala, and D. Lord (2007) Characterizing the Performance of a Bayesian Conway-Maxwell Poisson GLM. Accepted for presentation at the International Society for Bayesian Analysis, 9th World Meeting, Australia, 2008. To be submitted for publication in Computational Statistics and Data Analysis.



THANK YOU

VITA

Name: Srinivas Reddy Geedipally

Address: 309C, CE/TTI Bldg., Texas A&M University,
College Station, TX 77843-3136

Email Address: gsrinivas8@gmail.com

Education: B.E., Civil Engineering, Osmania University, 2002
M.Sc., Traffic Environment and Safety Management,
Linköpings University, 2005
Ph.D., Civil Engineering, Texas A&M University, 2008

Research Interests: Traffic Safety, Statistical Analysis of Crash Data, Bayesian Statistics, Crash risk, Geometric design, Traffic Flow Theory

Work Experience: Graduate Research Assistant, Texas Transportation Institute, Texas A&M University, 2005- 2008

Awards and Honors: Academic Excellence Award, Texas A&M University (2008-09).
Second prize winner, Student Research week, TAMU (2008).
Member, Pinnacle Honor Society (2008-).