EXAMINING THE POISSON-WEIBULL GENERALIZED MODEL FOR ANALYZING CRASH DATA

A Thesis

by

LINGZI CHENG

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

MASTER OF SCIENCE

August 2012

Major Subject: Civil Engineering

Examining the Poisson-Weibull Generalized Model for Analyzing Crash Data

EXAMINING THE POISSON-WEIBULL GENERALIZATION MODEL FOR ANALYZING CRASH DATA

A Thesis

by

LINGZI CHENG

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

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Dominique Lord
Committee Members,	Yunlong Zhang
	F. Michael Speed
Head of Department,	John Niedzwecki

August 2012

Major Subject: Civil Engineering

ABSTRACT

Examining the Poisson-Weibull Generalized Model
for Analyzing Crash Data. (August 2012)

Lingzi Cheng, B.S., Wuhan University of Technology
Chair of Advisory Committee: Dr. Dominique Lord

Over the last 20 to 30 years, there have been a significant number of statistical methods proposed for analyzing crash data. Traffic crashes are characterized as random and independent discrete non-negative events. Crash data have often been shown to exhibit over-dispersion. Therefore, the Negative Binomial (NB) is the preferred and widely used model to analyze this kind of data. Although NB model is very popular in traffic safety area, it still has limitations modeling crash data especially when crash data are characterized by low sample mean and small sample size. The main research objective of this thesis is to develop a new statistical method namely, Poisson-Weibull (PW) Generalized Linear Model (GLM) to analyze vehicle crash data and to evaluate its modeling performance at different dispersion levels. This study makes use of both simulated and observed data for accomplishing the research objectives.

The PW model is the mixture of Poisson and Weibull distributions. In this research, the statistical characteristics of the PW model were well defined and the parameters were estimated using a Bayesian approach. The PW model was initially evaluated using a series of simulated data for different dispersion levels. It was found that the PW model was able to reproduce and capture the true parameter values with high accuracy. After the initial analysis using the simulated data, the PW GLM was applied to two observed datasets and compared with the NB model. The goodness-of-fit (GOF) tests and model comparisons showed that the PW model performed as well as the NB model. Therefore, the PW model can be considered as an innovative and promising alternative for analyzing crash data.

DEDICATION

To my parents

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank those people who helped me to make this possible. First, I would like to thank my advisor, Dr. Dominique Lord. He was always ready to help when I confronted problems and he provided the observed data in this thesis. Second, I would like to thank my committee members, Dr. Michael Speed and Dr. Yunlong Zhang. They gave me valuable advice and suggestions about my thesis.

Thanks also go to the Ph.D. students of Dr. Lord, Srinivas Reddy Geedipally and Pei-Fen Kuo. It was Srinivas who brought the initial idea to me and provided help when I was doing my research. Pei-Fen was always nice and ready to help anyone. With encouragement and inspiration from her, I finally finished this. I would also like to thank all my fellow classmates and friends; they cheered me up and did not let me feel lonely when studying in the U.S.

Lastly, but most importantly, I would like to extend my special thanks to all my family, especially my parents. Being the only child of my family, studying abroad was a tough decision for me, as well as for my parents. But they are always there caring about me and supporting me both financially and emotionally.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Over the last 20 to 30 years, there have been a significant number of statistical methods proposed for analyzing crash data. Those models can be used for various purposes, such as establishing a relation between motor vehicle crashes and various covariates, predicting crash counts and screening variables. Traffic crashes are often characterized as random and are (assumed) independent discrete non-negative events, the most common probabilistic structures used for modeling crash data are the traditional Poisson and Negative Binomial (NB) distributions (Lord and Mannering 2010).

One significant characteristic of the crash data is over-dispersion, which means that the variance is greater than the mean. To accommodate the over-dispersion, the NB is the mostly preferred distribution because it captures the extra dispersion. In addition, its final equation has a closed form and the mathematics to manipulate the relationship between mean and variance is relatively simple (Hauer 1997).

Over the last few years, several studies have documented the limitations associated with the NB models. For example, the NB models have problems handling a large amount of zeros and a long or heavy tail (which creates highly dispersed data) (Shankar et al. 1997; Kumara and Chin 2003; Shankar et al. 2003); and when the sample mean becomes small, the traditional methods used to assess the goodness-of-fit (GOF) of general linear models (GLMs) can be highly unreliable and biased especially when maximum likelihood estimation (MLE) approach is used (Maycock and Hall 1984;

This thesis follows the style of *Journal of Transportation Engineering.*

Maher and Summersgill 1996; Wood 2002; Lord 2006). Given the limitations above, several alternative models have been examined. Lord and Mannering (2010) have documented these models in their recent study.

This thesis intends to introduce a new Poisson-based model, namely Poisson-Weibull (PW) generalized linear model (GLM) into the traffic safety literature, and evaluate its performance for modeling vehicle crash data at different dispersion levels.

1.1 Problem Statement

As the traditional regression models used for analyzing traffic crash data have their limitations, a new statistical model is needed to overcome those problems. The PW distribution is a mixture of the Poisson and Weibull distributions. As the mixing component (i.e., the Weibull distribution) has a large variety of shapes and scales, the PW model has the potential to overcome the limitations of the traditionally used models. This thesis will introduce the PW model into the traffic safety literature and evaluate its modeling performance at different dispersion levels. The modeling performance of the PW GLMs will be compared with the most commonly used NB model. If the PW model offers a potential for modeling traffic crashes and performs as well as the NB model, it will become another promising alternative in analyzing crash count data.

1.2 Research Objectives

The goal of this research is to develop the PW model to analyze vehicle crash data and evaluate its modeling performance at different dispersion levels. Several sub-objectives have been outlined to achieve this goal:

- document the statistical characteristics of the PW model, including the derivation of its probabilistic structure;
- estimate the parameters of the PW model; and,
- using the appropriate GOF measurements, assess the performance of PW GLM by comparing it with the NB model for the simulated and observed data at different dispersion levels

1.3 Thesis Organization

This thesis is divided into six sections. Section 1 provides an introduction to the thesis and problem statement, and the research objectives are included. Section 2 describes the background related to the crash data and statistical methods. It documents previous studies related to crash characteristics, key existing models and parameter estimation methods. Section 3 discusses the development of the PW model. Statistical derivation, characteristics and parameterization of the PW model are discussed in this section. Section 4 documents the initial assessment of the PW GLMs at different dispersion levels using the simulation data. Simulation protocol of data generation, parameter estimation results and modeling performance are discussed in this section. Section 5 describes the evaluation procedures of the PW GLMs using the observed data and documents its modeling performance by comparing the PW GLMs with the NB GLMs. Section 6 summarizes the tasks that are accomplished in this research and provides the concluding thoughts along with some avenues for the future study.

2. BACKGROUND

This section provides the basic background and insights into the area related to the model development and evaluation. The key component is to review the existing models that are used for analyzing crash data. The section's outline is as follows: first, the characteristics of crash data are discussed. Second, the statistical modeling methods for analyzing crash data are reviewed and the most commonly used models are discussed in detail. Lastly, different parameterization approaches are documented.

2.1 Crash Data Characteristics

To examine the relation between influential variables and vehicle crashes, safety analysts have used various statistical methods to model crash data. Important data and methodological issues have also been discussed over the years because those issues might be a potential source of error in terms of incorrect model specification.

2.1.1 Dispersion

Crash data mostly exhibit over-dispersion and rarely under-dispersion. As documented by Lord and Mannering (2010), the most notable characteristic of crash counts is over-dispersion, which means the variance is larger than the mean of the crash counts. The fundamental explanation for over-dispersion, explained by Lord et al. (2005) is that a traffic crash can be viewed as the product of Bernoulli trials with unequal probabilities (also known as Poisson trials). Thus, if a Poisson regression model, where the mean and variance are assumed to be equal, is applied when the data exhibits over-dispersion, the parameter estimates might be biased and the inference could be erroneous (Maycock and Hall 1984).

Though it is rare, crash data may also exhibit under-dispersion, the property where the variance is smaller than the mean of crash counts. It is found that many traditional count-data models produce incorrect parameter estimates (Lord and Mannering 2010; Oh et al. 2006).

2.1.2 Small-sample-size and Low-sample-mean

Crash data are often characterized by a limited sample size. Thus, the large-sample properties of some parameterization methods (e.g., maximum likelihood estimation) are not applicable. The estimates from the small sample size might be biased and lead to erroneous inferences (Lord and Mannering 2010). Also, when crash data are characterized as low-sample-mean, the distribution of the crash counts will be highly skewed towards to zero. In this way, it would also negatively affect the accuracy of parameter estimates and inferences (Maycock and Hall 1984; Maher and Summersgill 1996; Lord and Bonneson 2005). In general, the modeling complexity becomes significant when crash counts are characterized with small-sample-size or low-sample-mean.

2.1.3 Temporal and Spatial Correlation

Crash data are temporally and spatially correlated. In general, over a certain period of time, some explanatory variables might change. Ignoring the potential variation may result in the loss of important information in identifying the causal relation between independent variables and the crash counts (Lord and Mannering 2010; Washington et al. 2009). To avoid losing such information, crash data are usually divided into small time intervals. However, it means the same roadway entity will generate multiple observations with other unobserved effects remaining the same. Therefore, all these observations are correlated over time. From a statistical perspective, this correlation would negatively affect the model estimation. In the similar way, the spatial correlation also exists because the roadway entities may also share some unobserved effects (Lord and Mannering 2010; Washington et al. 2003; Shankar et al. 1998; Lord and Persaud 2000; Washington et al. 2009).

2.1.4 Injury Severity and Collision Type

Crash data are classified according to the injury severity or collision type. Injury severity can be classified as fatal, incapacitating injury, non-incapacitating injury, possible injury or no injury. Also, a collision type can be divided into rear-end, single-vehicle run-off-the-road, right-angle, and sideswipe among others. Commonly, injury severities or collision types are determined after the overall crash counts are obtained. If injury severities or collision types are modeled separately, consideration should be given to the correlation issue (Carson and Mannering 2001; Lee and Mannering 2002; Miaou and Song 2005; Park and Lord 2007; Geedipally and Lord 2010).

2.1.5 Under-reporting

Another important characteristic is that crash data are often under-reported. Elvik and Mysen (1999) stated that incomplete crash reports had been a major problem in highway safety analysis. Some studies have indicated that fatal crashes are most likely to be reported while no-injury crashes are most likely under-reported. The reporting threshold of a crash also depends on reporting agencies (Aptel et al. 1999; Hauer and Hakkert 1988).

2.2 Functional Form for Modeling Crash Data

To model the crash mean and establish a relation between crashes and explanatory variables, several functional forms can be used. The function form differs based on roadway entity types; typically a roadway entity can be categorized as an intersection or a segment. The flow-only functional forms are often used for analyzing intersection crashes. Although it is not considered as the most adequate functional form, it under-perform near the boundary condition (at least for intersection), it is still relevant for analyzing highway safety (Lord et al. 2008b). The most commonly used functional form for segments without covariates is given as follows:

$$\mu_i = \beta_0 \times L_i \times F_i^{\beta_1} \tag{2.1}$$

In a full model, site specific covariates need to be added in intersection models (Miaou and Lord 2003; Geedipally 2009). And when the covariates are included in the model, the functional form for roadway segments is:

$$\mu_i = \beta_0 \times L_i \times F_i^{\beta_1} \times e^{\sum_{j=2}^{J} X_{ij}\beta_j} \tag{2.2}$$

where μ_i is the estimated number of crashes for site i; L_i is the length of a segment i; F_i is the average annual daily traffic (AADT) for site i; $X_{ij}s$ is a series of covariates (e.g., shoulder width, lane width, etc.) for site i; and βs are the estimated coefficients.

2.3 Review of Existing Models

A variety of statistical modeling methods have been proposed and applied in traffic safety. Given the characteristics of crash data described above, each method has its strengths and limitations. The following section gives detailed description about commonly used modeling methods.

2.3.1 Poisson Regression Model

Since traffic crash is a discrete and non-negative event, the start point of modeling is to use a Poisson regression model. In the Poisson regression model, road entity i having Y_i crashes in a certain time interval is structured as:

$$Y_i \sim Poisson(\mu_i) \tag{2.3}$$

And the probability of roadway entity i having Y_i crashes is given by:

$$P(Y_i) = \frac{\exp(-\mu_i)\mu_i^{Y_i}}{Y_i!} \tag{2.4}$$

where μ_i is the Poisson mean and equals to roadway entity i's expected number of crashes $E(Y_i)$. Usually μ_i is specified as a function of explanatory variables and it is structured as:

$$\mu_i = \exp(\mathbf{\beta X_i}) \tag{2.5}$$

where $\mathbf{X_i}$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of parameters to be estimated.

As mentioned above, the Poisson regression model requires the mean and variance of crash counts to be equal. Therefore, when crash data exhibit the characteristic like over-dispersion or under-dispersion, the estimates and inferences might be problematic. At the same time, low-sample-mean and small-sample-size issues would also negatively affect the model precision.

2.3.2 Negative Binomial Regression Model

To deal with the over-dispersion situation of crash data, the Negative Binomial (NB) model has been widely used. It is also known as the Poisson-gamma distribution since it is structured as a mixture of the Poisson and gamma distribution. The NB model can be derived by structuring the Poisson mean as:

$$\mu_i = \exp(\boldsymbol{\beta}\mathbf{X_i} + \varepsilon_i) \tag{2.6}$$

where $\exp(\varepsilon_i)$ is assumed to be gamma distributed with mean equal to 1 and variance equal to α (also known as dispersion parameter). The crash variance, based on the NB model, can be defined as:

$$Var(Y_i) = E(Y_i)[1+\alpha E(Y_i)] = E(Y_i) + \alpha E(Y_i)^2 \tag{2.7}$$

where $E(Y_i)$ and $Var(Y_i)$ are the mean and variance of the crash count at site i respectively; and α is the dispersion parameter. Based on those assumptions, the probability mass function (p.m.f.) of the NB model is given as:

$$P(Y = Y_i; \mu, \varphi) = \frac{\Gamma(\varphi + Y_i)}{\Gamma(\varphi)\Gamma(Y_i + 1)}\left(\frac{\varphi}{\mu+\varphi}\right)^{\varphi}\left(\frac{\mu}{\mu+\varphi}\right)^{Y_i} \tag{2.8}$$

where $E(Y_i) = \mu$ is the mean of crash counts and $\varphi = 1/\alpha$ is the inverse dispersion parameter; and the variance of the crash count is given by:

$$Var(Y_i) = \frac{1}{\varphi}\mu^2 + \mu \tag{2.9}$$

When the dispersion parameter α approaches zero, the Poisson model is a limiting model of the NB model. Therefore the selection between those two models depends on the dispersion level of crash counts (Lord and Mannering 2010).

However, the NB model has a problem dealing with the data exhibiting under-dispersion. Also, when the sample size is small or the sample mean is low, there might be problems associated with the parameter estimation (Lord 2006; Lord and Mahlawat 2009).

2.3.3 Zero-inflated Poisson and Negative Binomial Models

Sometimes crash data have more zeros than expected under the assumption that they are Poisson or NB distributed. Therefore to accommodate the crash data that have preponderance of zeros, zero-inflated Poisson (ZIP) and zero-inflated NB (ZINB) models have been proposed. The Zero-inflated (ZI) models assume that the excess zeros come from two regimes or two distinct distributions. In other words, crash data are generated through a dual-state process (Cohen 1963; Rider 1961). The underlying assumption of ZI models when applied in traffic safety is that roadway entities exist in two states: a true-zero state and a non-zero state. A true-zero or inherently safe state is also known as virtually safe state to avoid having to defend the notion that roadway sites can be perfectly safe. In this state, it is believed that no crash data have been recorded or the probability of crash occurrence is extremely low. In an imperfect state, the non-zero state, it is assumed that crash occurrence follows a Poisson or a NB distribution (this state also has zero count sites). The probability of a roadway entity being in a zero or non-zero state can be further determined by a binary logit or probit model (Washington et al. 2009; Lambert 1992). Based on the Vuong statistic, many transportation safety analysts believed that ZIP and ZINB could provide a better fit than the Poisson and NB models for the given data (Shankar et al. 1997; Kumara and Chin 2003; Shankar et al. 2003; Lee and Mannering 2002).

However, Lord et al. (2005) in their researches argued that ZI models do not provide a defensible approach for modeling vehicle crashes, even when crash data has a large number of zeros. First, from a logic perspective, one should never claim a roadway entity to be safe but claim it either being more or less safer than the other one. In other words, the safety performance of a roadway entity should be stated in a relative term. Second, the true-zero state of ZI models has a long-term mean equal to zero, which also cannot properly reflect the crash-data generating nature. Lastly, it is ambiguous to define the boundary between those two states.

2.3.4 Conway-Maxwell-Poisson Model

The Conway-Maxwell-Poisson (COM-Poisson) distribution is a generalization of the Poisson distribution introduced by Conway and Maxwell (1962) for modeling queues and service rates. Based on Conway and Maxwell's work, Shmueli et al. (2005) further explored the statistical properties of the COM-Poisson model. Additionally, the conjugate distribution for the parameters had been developed by Kadane et al. (2006). The p.m.f. of the COM-Poisson model is given as follows:

$$P(Y = Y_i) = \frac{1}{Z(\lambda, v)} \frac{\lambda^y}{(y!)^v} \qquad (2.10)$$

$$Z(\lambda, v) = \sum_{n=0}^{\infty} \frac{\lambda^n}{(n!)^v} \qquad (2.11)$$

where Y_i is a discrete count; λ is a centering parameter that is approximately the mean of the observations; and v is the shape parameter of the COM-Poisson distribution.

The most notable characteristic of the COM-Poisson distribution is that it can accommodate both over-dispersion ($v < 1$) and under-dispersion ($v > 1$) situations. It can be approved that several common probability density distributions like the geometric distribution, the Bernoulli distribution and the Poisson distribution are special cases of the COM-Poisson distribution. In this sense, the flexibility of the COM-Poisson model to model different types of crash data has been greatly expanded.

However, as the likelihood function of the COM-Poisson distribution does not have a close form, the parameterization is complex. The modeling performance of the COM-Poisson distribution is also adversely affected when small-sample-size and low-sample-mean issues occur (Lord and Mannering 2010; Lord et al. 2008b).

2.3.5 Other Models

Over the last 20 to 30 years, there have been a significant number of statistical models proposed for analyzing crash data. Apart from the models that discussed above, there are many other models such as gamma model, generalized estimating equation model, generalized additive models, random-effects models, random-parameter models, bivariate/multivariate models, duration models, hierarchical/multilevel models, Bayesian neural network and support vector machine models that are not covered here. Interested readers can refer to Lord and Mannering's review paper (Lord and Mannering 2010).

2.4 Parameterization Methods

No matter what statistical model is used to analyze crash data, the parameters involved need to be estimated. Generally, maximum likelihood estimation (MLE) and Bayesian method are the two most commonly used techniques for estimating the parameters.

MLE selects the values for unknown parameters that would maximize the likelihood function (Casella and Berger 2001). For crash data analysis, the main advantage of MLE is that close-form functions often exist for most distributions.

However, for some complicated problems, MLE technique cannot be used, especially when the likelihood function is difficult to characterize (Lord and Mannering 2010).

Another way of parameter estimation is to use Bayesian methods. Compared to the MLE, the Bayesian methods can deal with complex models, especially when the likelihood function cannot be easily derived. The parameter estimates and inferences can be obtained by using Markov Chain Monte Carlo (MCMC) sampling method. The MCMC is the Monte Carlo integration using Markov chain. The Monte Carlo integration draws samples from the target distribution and then forms sample averages to estimate expectations (Gilks et al. 1996). However, due to the large computation involved, the simulation time of running Bayesian estimating process is significantly high. And the computational time increases as the data size and model complexity increase (Lord and Mannering 2010).

2.5 Summary

This section has provided the background of this thesis and summarized the previous studies related to traffic crash characteristics, functional forms, existing statistical models, as well as parameter estimation methods. Important data and methodological issues have been discussed in this section because those issues might be a potential source of error in terms of incorrect model specification. A brief review about existing statistical models is presented in this section. The last section has introduced two most commonly used parameter estimations methods, the MLE and Bayesian methods. The next section describes the characteristics and properties of the PW model.

3. STATISTICAL PROPERTIES OF THE POISSON-WEIBULL MODEL

This section describes the statistical properties of the PW GLM. It is organized in the following ways: firstly, the Poisson and Weibull distributions are discussed respectively. Then the PW GLM is derived. The characteristics of the PW GLM and its estimation approach are also discussed in this section.

3.1 Poisson and Weibull Distributions

In this section, statistical characteristics of Poisson and Weibull distributions are discussed for the deviation of the PW model.

3.1.1 Poisson Distribution

The statistic definition for Poisson process can be defined as follows (Montgomery and Runger 2003): “Given an interval of real numbers, assume counts occur at random throughout the interval. If the interval can be partitioned into subintervals of small enough length such that: 1. The probability of more than one count in a subinterval is zero, 2. The probability of one count in a subinterval is the same for all subintervals and proportional to the length of the subinterval, and 3. The count in each subinterval is independent of other subintervals, the random experiment is called Poisson process.” Therefore, the Poisson distribution is to express the probability of a given number of events occurring in fixed interval of time or space if these events occur with a known average rate and independent from each other. In transportation realm, it is assumed that the roadway entity i having Y_i crashes follows Poisson distribution. The probability mass function (p.m.f.) is given in (2.4). The mean and variance of Poisson distribution are given as follows:

$$E(Y_i) = \mu_i \tag{3.1}$$

$$Var(Y_i) = \mu_i \tag{3.2}$$

3.1.2 Weibull Distribution

The Weibull distribution is a continuous probability function and it is often used to model the time until a failure of many different physical systems. Its probability density function (p.d.f.) is given as (Montgomery and Runger 2003):

$$f(x) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}\exp[-\left(\frac{x}{\lambda}\right)^{k}] \tag{3.3}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the Weibull distribution. Depending on different values for k, the shape of the Weibull distribution has a wide variety and can be similar to that of the gamma, gamma-like, exponential or approximate normal distributions. The mean and the variance of the Weibull distribution are given as:

$$E(x) = \lambda\Gamma(1+\frac{1}{k}) \tag{3.4}$$

$$Var(x) = \lambda^2\Gamma(1+\frac{2}{k}) - [\lambda\Gamma(1+\frac{1}{k})]^2 \tag{3.5}$$

3.2 Poisson-Weibull Model Derivation

PW model is a mixture of the Poisson and Weibull distributions and intends to model the crash data whose output is thought to be generated by two different underlying mechanisms or different populations. Similar to the derivation of the NB model, the PW model assumes the model error which is independent of all covariates is Weibull distributed. The assumption here intends to capture the extra variation that the traditional NB model cannot fully handle.

The PW is derived as follows: the number of crashes Y_i at roadway entity i is assumed to be Poisson distributed with mean μ_i and independent over all entities:

$$Y_i \mid \mu_i \sim Poisson(\mu_i) \quad i = 1, 2, 3...I \tag{3.6}$$

The Poisson mean μ_i is structured as:

$$\mu_i = \rho_i \varepsilon_i = f(\mathbf{X\beta}\ ) \cdot \varepsilon_i \tag{3.7}$$

And,

$$\rho_i = f(\mathbf{X;\beta}) = \exp(\beta_0 + \sum_{j=1}^{J} \beta_j X_j) \quad j = 1, 2, 3, \ldots J \tag{3.8}$$

$$\varepsilon_i \mid \lambda, k \sim Weibull(\lambda, k) \tag{3.9}$$

Where Xs are the independent variables; J represents the total number of independent variable; βs are the regression coefficients; and ε_i is the model error independent of all covariates; k and λ are the scale and shape parameters of the Weibull distribution respectively.

Given of the assumption above, the PW distribution is defined as a mixture of those two distributions such that:

$$P(Y = y; \mu, \lambda, k) = \int Poisson(y; \rho\varepsilon) Weibull(\varepsilon; \lambda, k) d\varepsilon \tag{3.10}$$

To derive the function for the mean and variance of the PW model, law of iterated expectation is applied. The simple version of the law of iterated expectations is:

$$E(Y) = E_X[E(Y \mid X)] \tag{3.11}$$

Similarly, the variance is given as:

$$Var(Y) = E_X[Var(Y \mid X)] + Var_X[E(Y \mid X)] \tag{3.12}$$

By applying the law of iterated expectations, the mean and variance of the PW model in (3.10) can be derived as follows:

$$E(Y) = \rho E(\varepsilon) = \rho \times \lambda \Gamma(1+\frac{1}{k}) \tag{3.13}$$

$$Var(Y) = \rho \times \lambda \Gamma(1+\frac{1}{k}) + \rho^2 \times \lambda^2 \Gamma(1+\frac{2}{k}) - \rho^2 \times [\lambda \Gamma(1+\frac{1}{k})]^2 \tag{3.14}$$

3.3 Parameter Estimation of the PW Model

To estimate the parameters involved in the PW model, its likelihood function needs to be defined. However, the integration of equation (3.10) does not result in a closed function. In this sense, the MLE parameterization approach cannot be applied. This problem can be solved by using a hierarchical representation of equation (3.10). As discussed above, the PW model is conditional on the site-specific error term ε which explains the additional heterogeneity. Therefore the PW model can also be written as:

$$\begin{aligned} &P(Y = y; \mu \mid \varepsilon) = Poisson(y; \rho\varepsilon) \\ &\varepsilon \sim Weibull(\varepsilon; \lambda, k) \end{aligned} \tag{3.15}$$

The formula above has a nice Bayesian interpretation and hierarchical structure. Therefore, the parameter estimation and inference can be obtained by using MCMC.

To conduct the MCMC sampling method, WinBUGS is used in this research (Spiegelhalter et al. 2002). WinBUGS is part of the BUGS (Bayesian inference Using

Gibbs Sampling) project, which is a flexible software for the Bayesian analysis of complex statistical models using MCMC methods. It is a free open resource on the internet. Researchers in the past had successfully used WinBUGS to obtain parameter estimates and inferences (see e.g. Geedipally et al. 2011).

To obtain reliable parameter estimates from WinBUGS, several setups for running the MCMC need to be defined.

3.3.1 Prior Distribution

Bayesian formulation requires priors for unknown parameters. Informative or non-informative priors can be used. Prior distribution, which considers parameters as random variables, is the major difference between classical statistical theory and Bayesian approach. The prior distribution indicates the information available to the researcher before any "data" have been involved for analysis. When the prior distribution is combined with likelihood, the posterior distribution can be obtained. The posterior distribution is the base for any statistical inference. Therefore, the prior distribution for each unknown parameter in the PW model needs to be defined initially.

Since there are no related references about the PW model application in crash analysis area and the author has no other available information about the model error term, non-informative priors are used for all the unknown parameters in the PW model.

3.3.2 Equilibrium

MCMC sample method is based on the construction of a Markov chain that eventually converges to the equilibrium of the posterior distribution. It is found that a very efficient tactic in practice is to run multiple chains with different starting points. When the lines of different chains mix or cross-in trace is observed, the convergence is ensured. Therefore, the proper number of Markov chains should also needs to be determined.

The convergence of Markov chains can be monitored not only by the trace plots but also by other convergence diagnostics, such as Gelman-Rubin statistics. In WinBUGS, running a MCMC simulation would automatically generate several diagnostics including trace plot, history plot and Gelman-Rubin (G-R) statistics. For model evaluation and comparison, it was suggested that convergence was achieved when the G-R statistic was less than 1.2 (Mitra and Washington 2007).

In this study, 3 Markov chains are implemented and trace plot, history plot and G-R statistic are the three diagnostics considered simultaneously to check the convergence.

3.4 Summary

This section has documented the development of the PW model. As a mixture of the Poisson and Weibull distributions, the PW model assumes the model error which is independent of all covariates is Weibull distributed. The Poisson and Weibull distributions are briefly discussed for the derivation of the PW model and then the statistical characteristics of the PW model have been well defined in this section. As the PW model has a nice Bayesian interpretation and hierarchical structure, the Bayesian method is used to estimate the parameters. The next section presents the modeling performance of the PW model using simulated data.

4. PERFORMANCE OF THE POISSON-WEIBULL GLM

This section describes the initial assessment of the PW model. The main focus is to examine the modeling performance at different dispersion levels. All the computations and interpretations in this section will be evaluated under the Bayesian settings.

Section 4 is organized in the following way: first, study methodology is provided and simulation and testing protocols are discussed. Second, simulation results are presented and computational analyses are conducted. Lastly, a brief discussion based on the results is provided.

4.1 Methodology

To initially evaluate the performance of the PW GLM, a number of datasets are simulated from the PW model. Predefined values for the parameters, also called as “true parameters,” are assigned corresponding to different dispersion levels. Then the parameters are re-estimated by using the MCMC approach. The estimates given by the MCMC are compared to the true parameters to examine the overall performance of the PW model.

4.1.1 Data Simulation Protocol

To generate the simulated data from the PW model, the simulation protocol is described as follows:

1) Generate a mean value (ρ_i) for roadway entity i from a fixed sample population mean (δ):

$$\rho_i = \delta$$

2) Generate a value (ε_i) from a Weibull distribution with two parameters k and λ. Here, the mean of the Weibull distribution is equal to 1. The values for k and λ are determined based on the dispersion parameter (α):

$$\varepsilon_i \sim Weibull(k, \lambda)$$

3) Calculate the mean (μ_i) for roadway entity i:

$$\mu_i = \rho_i \times \varepsilon_i$$

4) Generate a discrete value (Y_i) for entity i from a Poisson distribution with mean equal to μ_i:

$$Y_i \sim Poisson(\mu_i)$$

5) Repeat steps 1 to 4 N times for the number of observations which is also considered as the sample size.

The simulation process was conducted by using R with the following values:

1) Sample size "N": 300

2) Dispersion level "α": 0.5, 1, 2, 3,5

3) Sample population mean "δ": 10

To avoid small-sample-size and low-sample-mean influences on the initial evaluation, the values for N and δ were set at the moderate levels. Therefore, there were 5 scenarios in total used to assess the modeling performance of the PW model corresponding to the different dispersion levels. For each combination of sample size, dispersion level and sample population mean, the simulation was replicated 100 times.

The assigned values for k and λ based on the dispersion levels are given in the following table:

Table 4-1. Predefined Values for Weibull Parameters

Dispersion Parameter α	k	λ
0.5	1.436	1.101
1	1.000	1.000
2	0.721	1.096
3	0.607	1.118
5	0.500	0.500

Due to the different coding scheme in WinBUGS, $\omega = 1/\lambda^k$ was introduced where k and λ are the scale and shape parameters of the Weibull distribution in (3.3). Although the Weibull distribution was re-parameterized, it did not influence the model performance. Therefore, the predefined values for k and ω were given in Table 4-2:

Table 4-2. Predefined Values for Re-parameterized Weibull Distribution

Dispersion Parameter α	k	ω
0.5	1.436	0.871
1	1.000	1.000
2	0.721	0.936
3	0.607	0.935
5	0.500	1.414

4.1.2 Testing Protocol

In WinBUGS, non-informative priors for k and ω were used (i.e., gamma (0.1, 0.1) priors). A total of 3 Markov chains were selected with 50,000 iterations per chain. The first 25,000 iterations were discarded as burn-in samples and the rest were used to conduct the estimation and inference. The G-R statistic was used to ensure the convergence. When the G-R statistic was below 1.1, the convergence was considered to be achieved (Mitra and Washtington 2007).

4.2 Simulation Results

This part documents the estimated values for the parameters and computational analyses based on the simulated data. Corresponding discussions are also provided here.

4.2.1 Parameter Estimates

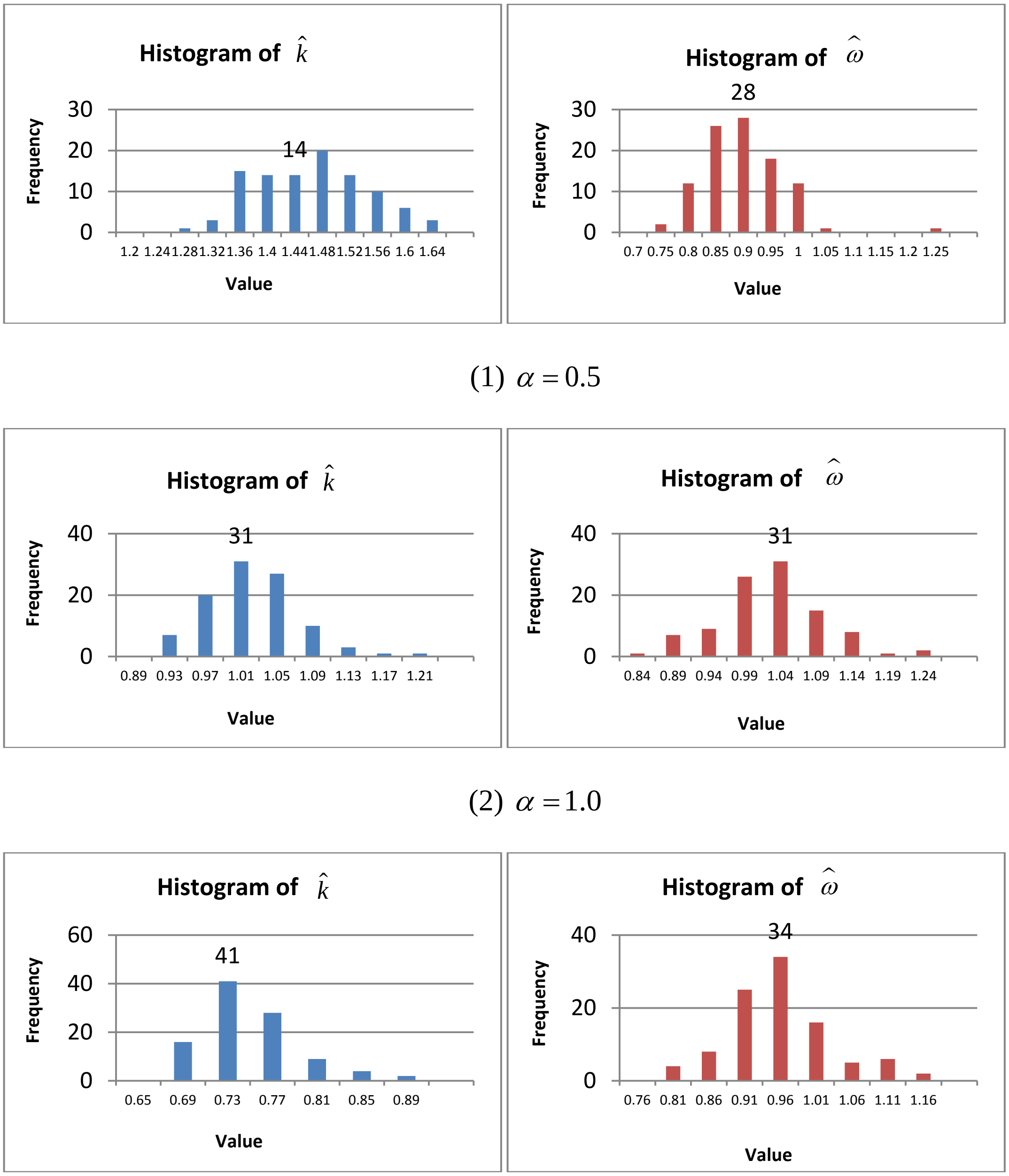
Since there were 100 simulation replications for each scenario, the average for each parameter was calculated. The following table summarizes the results for each dispersion level.

Table 4-3. Simulation Results

Dispersion Level	$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 2.0$		$\alpha = 3.0$		$\alpha = 5.0$	
Parameters	k	ω	k	ω	k	ω	k	ω	k	ω
Theoretical Values	1.436	0.871	1.000	1.000	0.721	0.936	0.607	0.935	0.500	1.414
Estimated Values	1.446	0.873	1.000	0.993	0.731	0.931	0.607	0.936	0.506	1.414
Standard Deviation	0.080	0.061	0.053	0.065	0.038	0.060	0.032	0.061	0.029	0.084
Min.	1.247	0.747	0.840	0.084	0.659	0.767	0.534	0.796	0.414	1.227
Max.	1.631	0.748	1.202	0.192	0.870	1.130	0.686	1.096	0.565	1.565

It can be seen in Table 4-3 the estimated value for each parameter is very close to the theoretical value. Besides, with the increase of dispersion level, it is found the standard deviation for $\hat{k}$ becomes smaller.

Fig. 4-1 illustrates the histograms for $\hat{k}$ and $\widehat{\omega}$ estimates for each dispersion level. The bin with a label on top indicates that the theoretical value is within this value range. This indicates that the PW model was able to reproduce the "true" parameters with certain accuracy. Additionally, it can be seen when the value of α decreases, the estimates for $\hat{k}$ are distributed more dispersedly while the estimates for $\widehat{\omega}$ do not fluctuate much for different dispersion levels. This might be explained by the unknown variation for the parameter λ since it was re-parameterized.

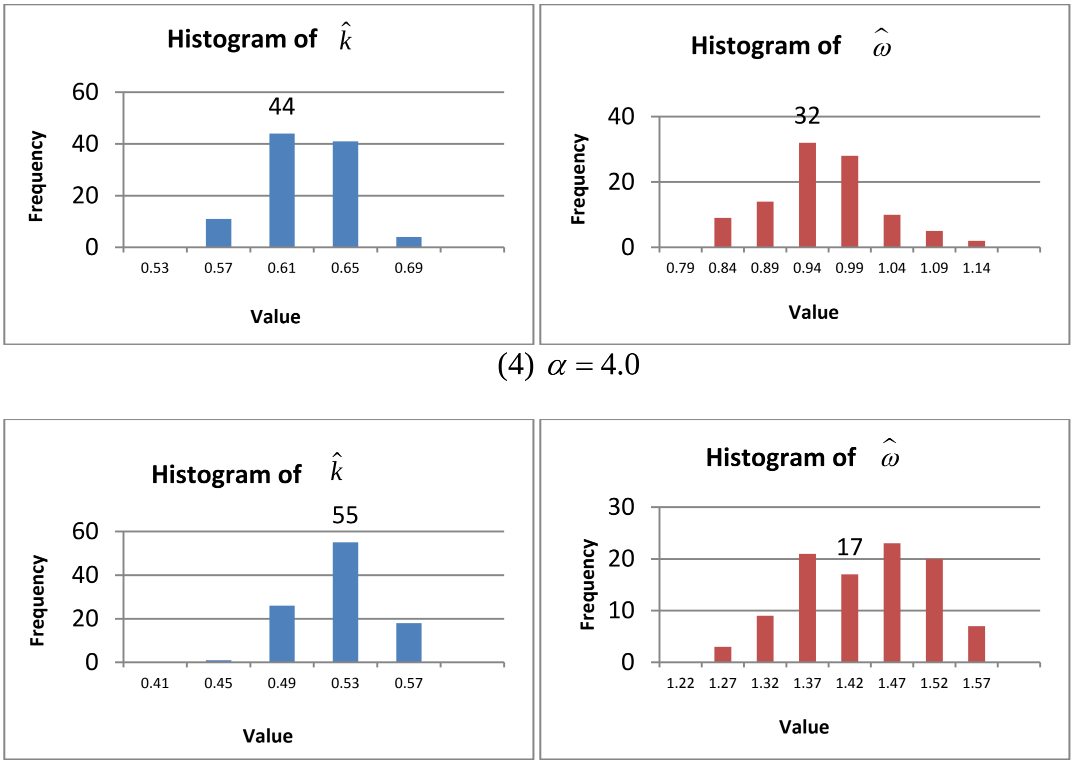


(1) $\alpha = 0.5$

(2) $\alpha = 1.0$

(3) $\alpha = 2.0$

Fig. 4-1. Histograms for $\hat{k}$ and $\hat{\omega}$ at Different Dispersion Levels



(4) $\alpha = 4.0$

(5) $\alpha = 5.0$

Fig. 4-1. Continued

WinBUGS displayed the median, 2.5th and 97.5th percentiles for the estimated parameter by default, it have been examined that all the estimated parameters lied in 95% credible intervals.

4.2.2 Computational Analyses

In order to examine the estimation accuracy of the PW model at different dispersion levels, three summary statistics were used.

4.2.2.1 Bias

The bias of an estimator is the difference between the estimator's expectation and the true value of the parameter that being estimated. Therefore, the bias can be defined as:

$$Bias = E(\hat{\theta}_r) - \theta_r \tag{4.1}$$

where r is the number of replication. For each replication, the bias was calculated. The bias summaries based on different dispersion levels are illustrated in Table 4-4 and Fig. 4-2, 4-3:

Table 4- 4. Bias Summaries at Different Dispersion Levels

Dispersion Level	$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 2.0$		$\alpha = 3.0$		$\alpha = 5.0$	
Estimated Parameters	$\hat{k}$	$\widehat{\omega}$	$\hat{k}$	$\widehat{\omega}$	$\hat{k}$	$\widehat{\omega}$	$\hat{k}$	$\widehat{\omega}$	$\hat{k}$	$\widehat{\omega}$
Average	0.010	0.002	0.004	0.000	0.010	-0.005	-0.001	0.001	0.006	0.000
Min.	-0.189	-0.124	-0.109	-0.160	-0.062	-0.169	-0.073	-0.139	-0.086	-0.187
Max.	0.195	0.379	0.202	0.192	0.149	0.194	0.079	0.161	0.065	0.151

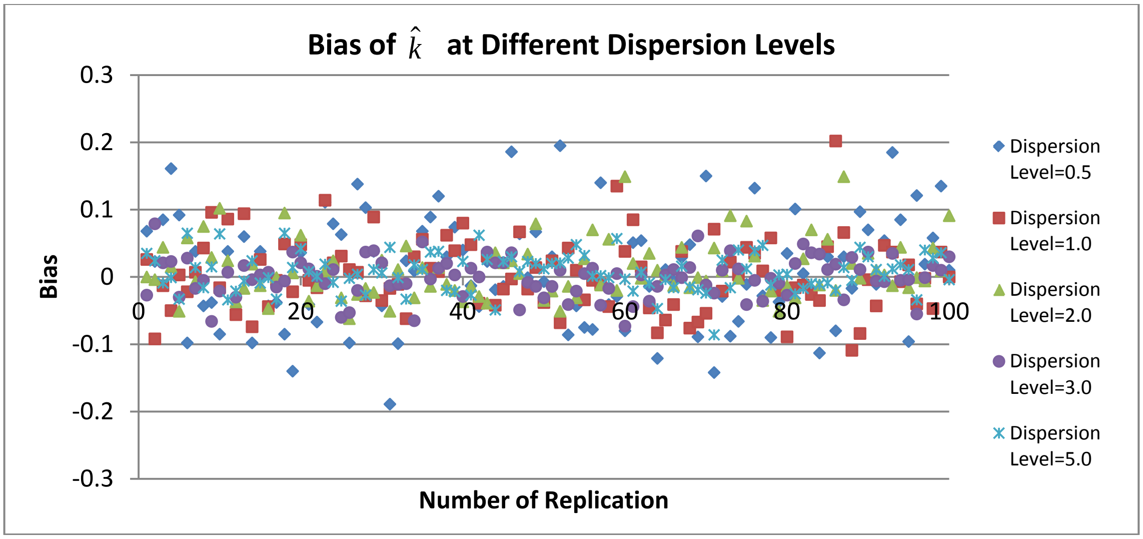


Fig. 4-2. Bias of $\hat{k}$ at Different Dispersion Levels

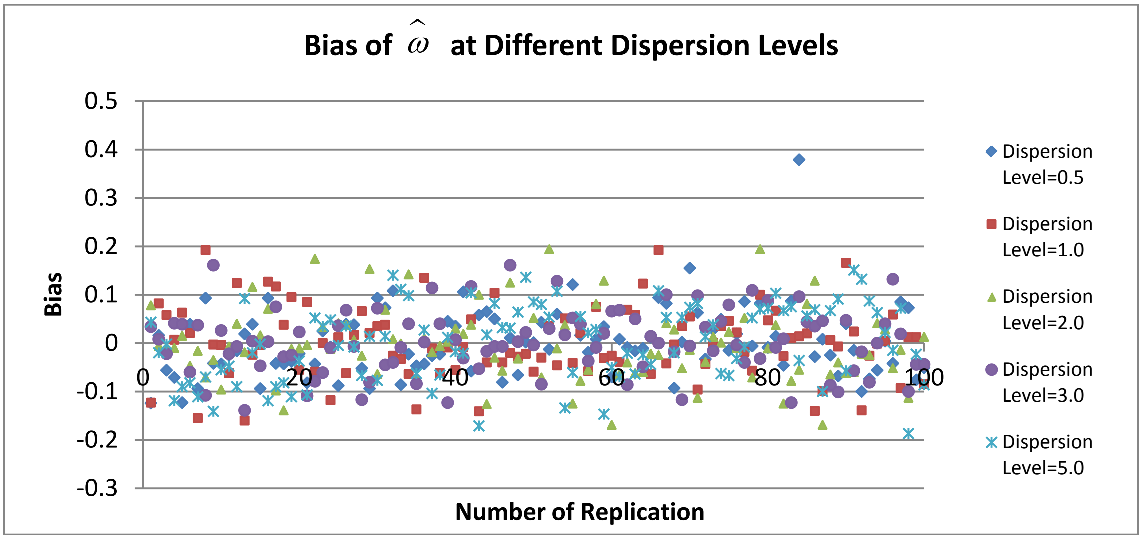


Fig. 4-3. Bias of $\hat{\omega}$ at Different Dispersion Level

By examining the Fig. 4-2 and Fig. 4-3 along with the results in Table 4-4, the average biases for parameters $\hat{k}$ and $\widehat{\omega}$ are small in regardless of the dispersion levels. And for parameter $\hat{k}$, the biases oscillated closer to the 0 for higher dispersion levels while for parameter $\widehat{\omega}$, there was no clear difference between each dispersion level.

4.2.2.2 Absolute Percent Difference (APD)

APD is calculated as:

$$APD = \frac{|\theta - \widehat{\theta}|}{\theta} \times 100\% \tag{4.2}$$

For each simulate replication, the APDs for $\hat{k}$ and $\widehat{\omega}$ were calculated. The summary statistics for 100 replications are provided in Table 4-5 and Fig. 4-4 and Fig. 4-5 document the APDs for each simulation replication.

Table 4-5. APD Summaries at Different Dispersion Levels

Dispersion Level	$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 2.0$		$\alpha = 3.0$		$\alpha = 5.0$	
Estimated Parameters	$\hat{k}$	$\widehat{\omega}$	$\hat{k}$	$\widehat{\omega}$	$\hat{k}$	$\widehat{\omega}$	$\hat{k}$	$\widehat{\omega}$	$\hat{k}$	$\widehat{\omega}$
Average (%)	4.653	6.075	4.155	5.640	4.405	5.935	3.743	5.499	4.458	4.662
Min. (%)	0.070	0.115	0.000	0.000	0.000	0.107	0.000	0.000	0.200	0.141
Max. (%)	13.579	43.513	20.200	19.200	20.666	20.727	13.015	17.219	17.200	13.225

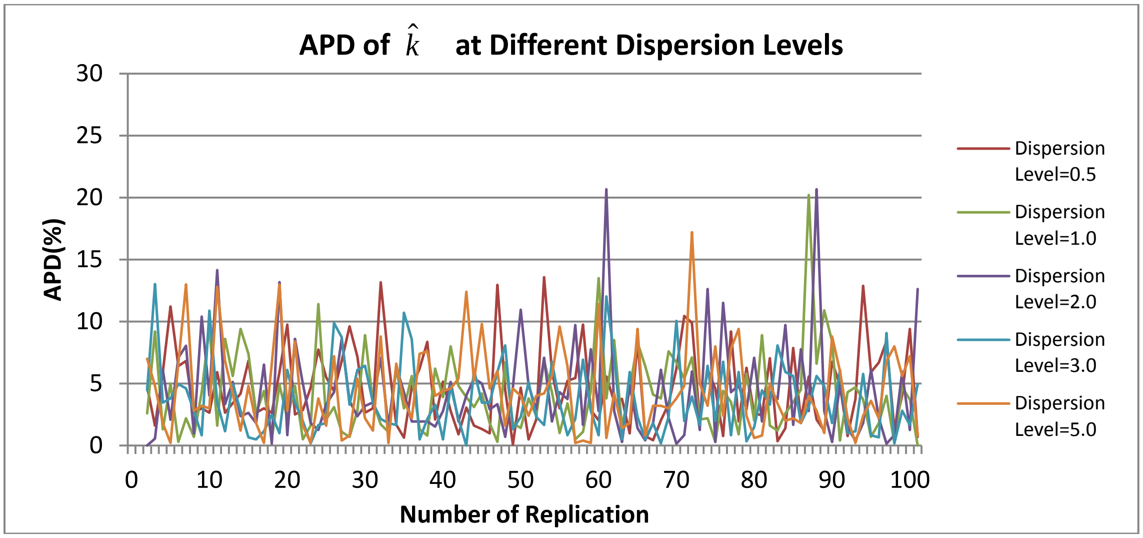


Fig. 4-4. APD of $\hat{k}$ at Different Dispersion Levels

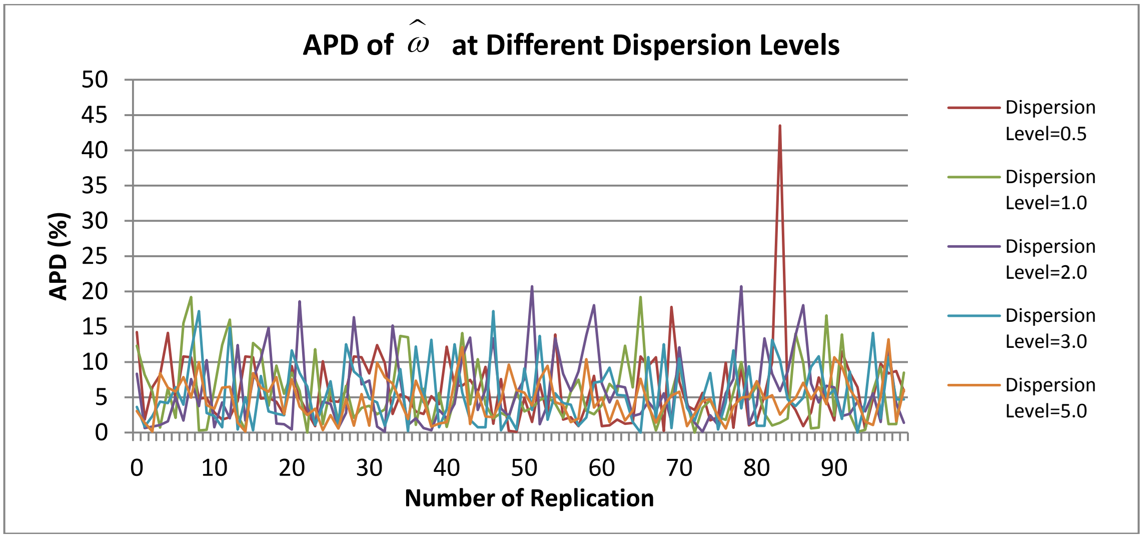


Fig. 4-5. APD of $\hat{\omega}$ at Different Dispersion Levels

Fig.4-4 and Fig.4-5 along with the results in Table 4-5 indicate that the average APD values for each dispersion level are small for both parameters. However, for parameter $\hat{\omega}$, it seems that for higher dispersion levels, the PW model has better performance based on APD values.

4.2.2.3 Root Mean Square Error (RMSE)

RMSE is also known as Root Mean Square Deviation (RMSD) and it is used to measure the difference between an estimator and the true parameter being estimated by taking account both bias and variance. RMSE is defined as:

$$RMSE = \sqrt{Bias^2 + Var} \tag{4.3}$$

where variance can be obtained from Table 4-3. Table 4-6 and Fig. 4-6 and 4-7 summarizes the RMSE for the parameters at each dispersion level:

Table 4-6. RMSE Summaries at Different Dispersion Levels

Dispersion Level	$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 2.0$		$\alpha = 3.0$		$\alpha = 5.0$	
Estimated Parameters	$\hat{k}$	$\hat{\omega}$	$\hat{k}$	$\hat{\omega}$	$\hat{k}$	$\hat{\omega}$	$\hat{k}$	$\hat{\omega}$	$\hat{k}$	$\hat{\omega}$
Average	0.109	0.085	0.071	0.092	0.053	0.088	0.041	0.085	0.038	0.111
Min.	0.076	0.059	0.050	0.061	0.035	0.059	0.030	0.058	0.027	0.082
Max.	0.218	0.388	0.212	0.206	0.156	0.206	0.086	0.175	0.090	0.202

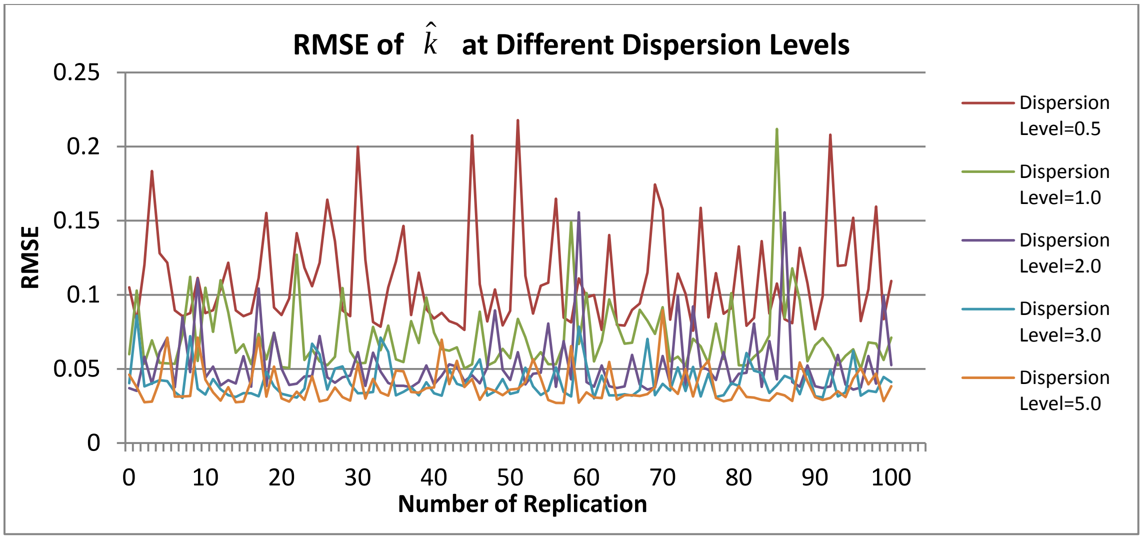


Fig. 4-6. RMSE of $\hat{k}$ at Different Dispersion Levels

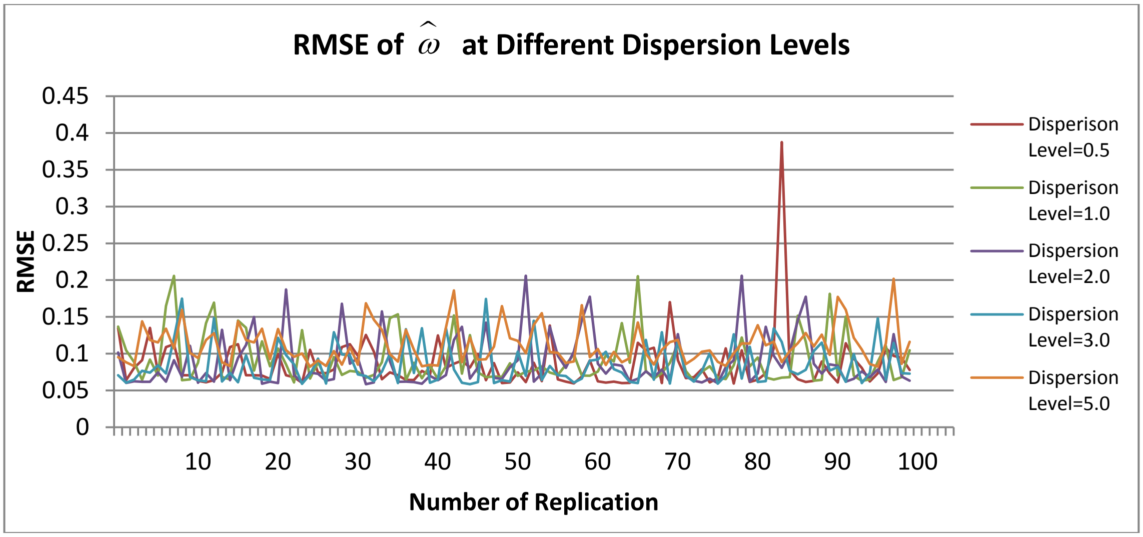


Fig. 4-7. RMSE of $\hat{\omega}$ at Different Dispersion Levels

Fig. 4-6 gives a clear sign that the estimation accuracy for $\hat{k}$ became more precise with the increase of the dispersion level. However, the RMSEs for $\hat{\omega}$ do not vary much as a function of different dispersion levels.

4.3 Discussions

Based on the simulation results above, it indicates that the PW GLM has a potential for modeling crash data. The following discussions are drawn from those analyses:

1) The predefined “true” values for the parameters lie in the 95% credible interval for all those 5 scenarios. In other words, the PW model was able to reproduce the “true” parameters. And the credible interval for parameter k tends to become narrower with the increase of the dispersion level. For parameter ω, even though it was re-parameterized with two parameters k and λ, the credible intervals do not change too much irrespective of dispersion levels. Moreover, when the value of α decreases, the estimates for $\hat{k}$ are distributed more dispersedly.
2) Based on the summary statistics of bias and APD, it shows that the PW model has the ability to capture the true value with considerably high accuracy regardless of different dispersion levels.
3) Taking bias and variance into account, the RSME results indicate that the estimation accuracy for parameter k becomes more precise with the increase of dispersion level while the averages for ω do not fluctuate too much. The RMSE plot for $\hat{k}$ has a clear sign indicating that the PW model has a better modeling performance to capture the true value for the parameter.

In general, the results based on the simulated data have shown that the PW model has the ability to reproduce and capture the true parameters with high accuracy. There is no sign that the modeling performance has been negatively affected by the dispersion levels even though the PW model seems to have a better modeling performance in dealing with high level of dispersed data. Therefore, the PW model provides the potential for modeling crash counts.

4.4 Summary

This section has documented the modeling performance of the PW GLM using simulated data. Data simulation protocol was first introduced in this section. The values for the two parameters in the PW model were predefined corresponding to the different dispersion levels. The estimated values were compared to the predefined values and computational analyses were conducted in the second section. Based on the bias, APD and RMSE analyses, it showed that the PW GLM could reproduce and capture the true parameter with high accuracy. Additionally, it seemed that the PW GLM had better modeling performance for higher dispersion levels. The next section presents the application of the PW GLM on observed data.

5. APPLICATION OF THE POISSON-WEIBULL GLM TO OBSERVED DATA

Based on the simulation results in Section 4, the PW GLM shows the potential to model crash count data irrespective of dispersion levels. Thus, there is a need to evaluate the PW modeling performance on observed traffic crash data.

The main objective of this section is to compare the PW GLM with the most widely used NB model in modeling crash counts using observed crash data. This section is organized as follows: firstly, the functional forms that are used to model crash data and selected GOF measuring statistics are introduced in the methodology section. Secondly, descriptions about two observed crash datasets are provided. Then the modeling results and corresponding analyses are presented. Lastly, some concluding thoughts are discussed.

5.1 Methodology

This part describes the methodology that used to estimate and compare the modeling performance of two models. There were two observed datasets and for each dataset the PW GLM and NB GLM were used to estimate the parameters. The comparison was achieved based on the resultant estimates and 4 selected GOF statistics.

Two major functional forms have been discussed in Section 2. The general expression of the functional form selected for the observed datasets in this section is as follows:

$$\mu_i = \beta_0 \times L_i \times F_i^{\beta_1} \times e^{\sum_{j=2}^{J} X_{ij}\beta_j} \tag{5.1}$$

As each dataset contains its own covariates $X_{ij}s$ exact functional form for the dataset will be presented in the data description section.

After the estimated coefficients βs were obtained from WinBUGS, the expected crash counts μ_i for each site i were calculated by using spreadsheet. To compare the modeling performance of the PW and NB GLMs, the GOF tests were conducted. The selected GOF statistics are introduced as follows:

5.1.1 Deviance Information Criterion (DIC)

When Bayesian estimation method is applied, the DIC is often used as a GOF statistic. It is defined as:

$$DIC = \overline{D} + p_D \tag{5.2}$$

where $\overline{D} = -2\ln L$ represents the posterior mean of the deviance of the un-standardized model and L is the mean of the model log likelihood; $p_D = \overline{D} - D(y\,|\,\overline{\theta})$ represents the penalty for the number of effective model parameters where $D(y\,|\,\overline{\theta})$ is the point estimate of deviance for the posterior mean $\overline{\theta}$. A smaller DIC refers to a better fit to the data. In general, as a rule of thumb, differences in the values of DIC of more than 10 definitely rule out the model with a higher DIC and a difference in DIC between 5-10 can be considered substantial (Spiegelhalter et al. 2002). However, the DIC is dependent on the structure of the model. Even though NB model is equivalent to the Poisson-Gamma model, the DIC values for those two models are much different. Thus, the PG model is used for all the following analysis.

5.1.2 Mean Absolute Deviance (MAD)

MAD is a statistic used to assess how well the model fit to the data and provides a measure of the average mis-prediction of the model. The MAD is defined as:

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |\,\hat{y}_i - y_i\,| \qquad (5.3)$$

5.1.3 Mean Squared Predictive Error (MSPE)

MSPE is typically used to assess the error associated with a validation or external dataset. It can be computed as follows:

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2 \tag{5.4}$$

5.1.4 Pearson Chi-Square (χ^2)

Pearson χ^2 is often used to assess the overall fit of a model. This statistic follows the χ^2 distribution with $n-p$ degree of freedom where p is the number of model variables. This statistic is asymptotic to the χ^2 distribution for larger sample sizes and is computed as:

$$\chi^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{Var(y_i)} \tag{5.5}$$

where $Var(y_i)$ is the crash frequency at site i and for PW and NB models, the variances can be obtained by using (3.6) and (3.7).

5.2 Data Description

This section describes the characteristics of two datasets that were used for comparing the PW and NB GLMs.

The datasets used in this research have several potential crash related variables. Although a simplified model with few variables may be preferred, leaving out important explanatory variable would result in omitted variable bias in parameter estimation and provide erroneous inference and crash count prediction. Possible endogenous relation issue between dependent and response variables should also be ruled out. After examining several candidate datasets based on those two principles, the following two datasets were selected.

5.2.1 Indiana Data

Crash and traffic data at 338 rural interstate road segments were collected for a 5-year period (1995-1999) in the state of Indiana. This data have previously been used for developing a tobit regression model and a negative binomial-Lindley model (Washington et al. 2009; Geedipally et al. 2011; Anastasopoulos et al. 2008). This dataset contains no reported crashes for 120 out of 338 highway segments over the 5-year observational period. Table 5-1 summarizes the basic statistics of crash, road geometric and traffic variable used in this study. For a complete and detailed list of variables, the interested reader is referred to Washington et al. (2009).

Table 5- 1. Indiana Data Summary

Traffic Variable	Min.	Max.	Average (Std. Dev)	Total
Number of Crashes (5 years)	0	329	16.97(36.30)	5737
Average Daily Traffic(ADT) (veh/day)	9442	143422	30237.56(28776.43)	_
Minimum Friction of Road Segment (FR) (0-100 scale)	15.90	48.20	30.51(6.67)	_
Pavement Surface(PS) (Asphalt=1,Concrete=0)	0	1	0.77(0.42)	_
Median Width (MW) (ft)	16	194.7	66.98(34.71)	_
Median Barrier(MB) (Present=1,Absent=0)	0	1	0.16(0.37)	_
Interior Rumble Strips(IRS) (Present=1,Absent=0)	0	1	0.72(0.45)	_
Segment Length(SL)(miles)	0.009	11.53	0.89(1.48)	300.09

Based on the summary above, the functional formal used for modeling Indiana data is as follows:

$$\mu_i = \exp[\ln\beta_0 + \ln(SL)_i + \beta_1\ln(ADT)_i + \beta_2(FR)_i + \beta_3(PS)_i + \beta_4(MW)_i + \beta_5(MB)_i + \beta_6(IRS)_i] \qquad (5.6)$$

5.2.2 Texas Data

The second dataset contained crash data collected at 4-lane rural undivided in Texas. The data were provided by the Texas Department of Public Safety (DPS) and the Texas Department of Transportation (TxDOT) and were used for the National Cooperative Highway Research Project (NCHRP) 17-29: Methodology for Estimating the Safety Performance of Multilane Rural Highways (Lord et al. 2008a). Table 5-2 presents the summary statistics of the data.

Table 5- 2. Texas Data Summary

Traffic Variable	Min.	Max.	Average (Std. Dev)	Total
Number of Crashes (5 years)	0	97	2.84(5.69)	4253
Average Daily Traffic (ADT) (veh/day)	42	24800	6613.61(4010.01)	_
Lane Width(LW) (ft)	9.75	16.5	12.57(1.59)	_
Total Shoulder Width (SW) (ft)	0	40	9.96(8.02)	_
Curve Density(CD) (curves/mile)	0	18.07	1.43 (2.35)	_
Segment Length (L) (miles)	0.1	6.28	0.55(0.67)	830.49

Based on Texas data, the functional form used is illustrated as follows:

$$\mu_i = \exp[\ln \beta_0 + \ln(SL)_i + \beta_1 \ln(ADT)_i + \beta_2 (LW)_i + \beta_3 (SW)_i + \beta_4 (CD)_i] \qquad (5.7)$$

5.3 Results

This part presents the modeling results for the PW and NB GLMs. The parameter estimates and GOF statistics are illustrated in this section. Based on the resultant estimates, examinations about prediction accuracy, covariate sensitivity analysis, model-checking and relation between predicted crash variance and mean were also conducted in this part.

5.3.1 Indiana Data

5.3.1.1 Parameter Estimate

Table 5-3 summarizes the coefficient estimates of the PW and NB models for the Indiana data as follows:

Table 5- 3. Coefficient Estimates for Indiana Data

Variable	**NB**		**PW**	
	Value	S.E.	Value	S.E.
Intercept $(\ln \beta_0)$	-4.627	1.354	-4.022	1.377
ADT (β_1)	0.7029	0.1254	0.6428	0.1243
FR (β_2)	-0.02589	0.01048	-0.02713	0.01128
PS (β_3)	0.4226	0.1874	0.4267	0.1992
MW (β_4)	-0.005169	0.001906	-0.005489	0.002007
MB (β_5)	-3.035	0.3047	-2.99	0.3093
IRS (β_6)	-0.3901	0.1866	-0.4113	0.1987
ϕ	1.089	0.1392		
ω			0.9959	0.316
k			0.9805	0.07021

The segment length was treated as an offset and thus the crash frequency increases linearly with the increase in segment length. For both PW and NB models, the estimated coefficient for the traffic flow is less than 1. This indicates that with the increase in the traffic flow, the crash risk increases in a decreasing rate. It should be noted that the 95% marginal posterior credible intervals for each coefficient estimate did not include the origin. Furthermore, all the estimated coefficients between the two models have the same sign, and most of their values are very close. The standard errors for estimated coefficients of PW model are slightly larger than those of NB model. Though NB and PW model have the same hierarchical structure, there is an extra parameter in the PW model. This is because the shape and scale parameters are assumed to be the same in the NB model but not in the PW model.

5.3.1.2. Goodness-of-fit Statistics

Based on the coefficient estimates in Table 5-3, the predicted crash counts were calculated by using the function form. To compare the modeling performance of the PW and NB model, the GOF tests were conducted. Table 5-4 summaries the GOF statistics of the PW and NB model for Indiana data:

Table 5- 4. GOF Statistics for Indiana Data

GOF	NB	PW
DIC	1486.47	1450.67
Pearson χ^2	1009.406	1003.086
MAD	6.919	7.014
MSPE	209.909	231.476

Based on the GOF results in Table 5-4, it is found both the DIC and Pearson χ^2 statistics indicate that the PW model fits the data better than the NB model. However, the MSPE statistic is in favor of the NB model while the MAD values for both models are very close.

5.3.1.3 Prediction Accuracy

Fig. 5-1 illustrates the histograms of the observed crash counts and predicted crash counts for both PW and NB models. The PW model provides similar predictions as the NB model. And compared to the observed crash counts, both the PW and NB model could predict the observed crash counts with certain accuracy. However, it should be noted that this histogram plot could only provide a thorough examination about the modeling performance of the two models since it is based on the frequency distributions.

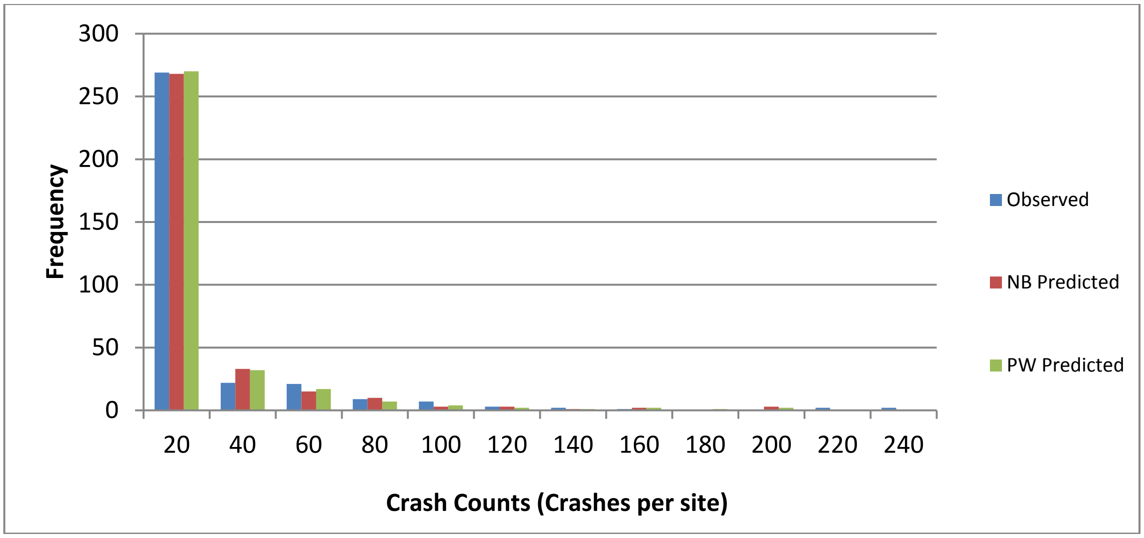


Fig. 5-1. Histograms of Observed and Predicted Crash Counts for Indiana Data

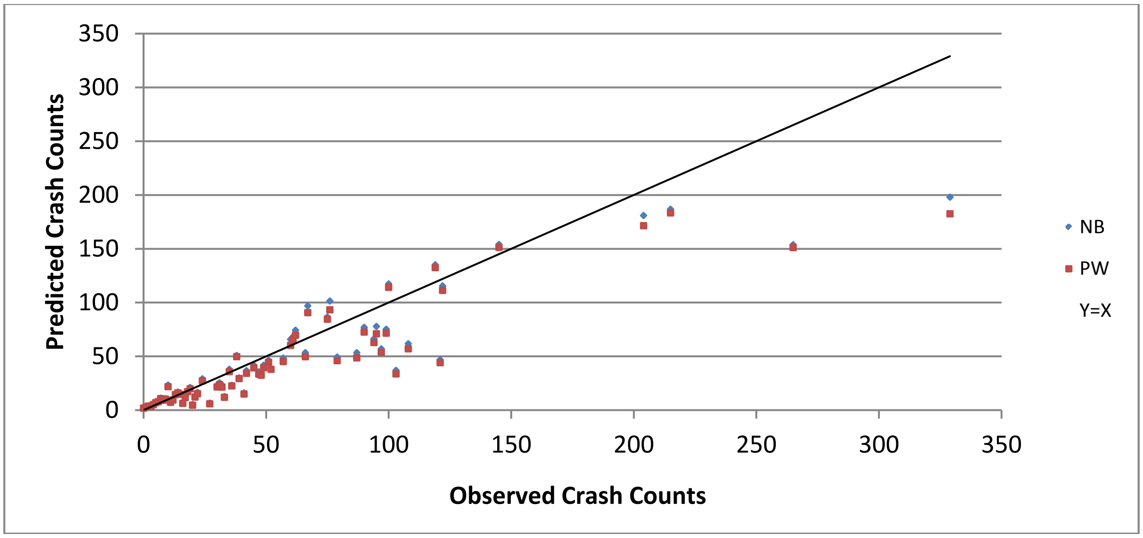


Fig. 5-2. Predicted Crash vs. Observed Crash for Indiana Data

Fig. 5-2 illustrates a more detailed comparison between predicted and observed crash counts. It can be seen that the PW model provides smaller predicted values than for the NB model. The average residual for the PW model is 2.19 and the minimum and maximum residuals are -29.56 and 146.48 respectively. For the NB model, the average, minimum and maximum residuals are 1.42, -37.53 and 131.01 respectively. Additionally when the observed crash counts are below 10 crashes/mile, both PW and NB models are over-estimating the crash counts. And when the observed crash counts are larger than 200 crashes/mile, both PW and NB models are under-estimating the crashes. The detailed table is presented in Appendix A.

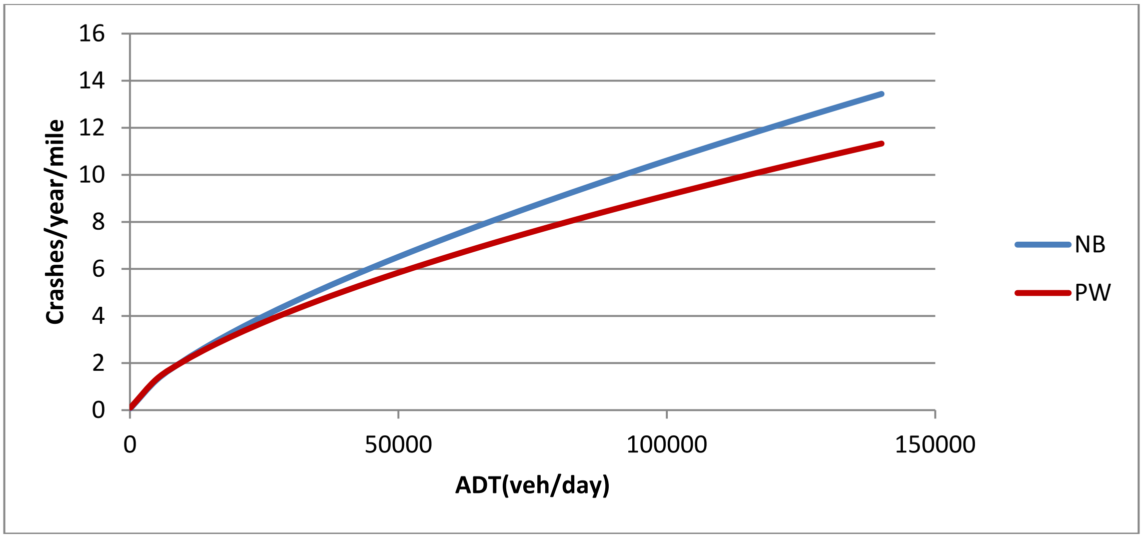


Fig. 5-3. Predicted Crash Counts vs. ADT for Indiana Data

Fig. 5-3 illustrates the comparison between the predicted crash counts per year per mile against the covariate ADT for Indiana data. It is found that the PW model provides lower estimates than the NB model except when the ADT flows are extremely low. Similar analysis can be conducted for other covariates and the figures are presented in the Appendix B.

5.3.1.4 Covariate Sensitivity

The functional form for modeling Indiana data is illustrated in equation (5.6). Therefore the sensitivity analysis for covariate ADT can be examined by using the following equation:

$$\frac{\partial \log(\mu)}{\partial ADT} = \beta_1 \frac{1}{ADT_i} \tag{5.8}$$

Fig. 5-4 shows the sensitivity of covariate ADT for both NB and PW models. It can be seen that those two models have quite similar trend. When the ADT is below 20000 veh/day, one unit increase in ADT would result in huge decrease in logarithm form of the estimated crash. However, this decrease becomes smaller when the ADT is at higher levels. Similar analysis can be conducted for other covariates and the figures are presented in the Appendix C.

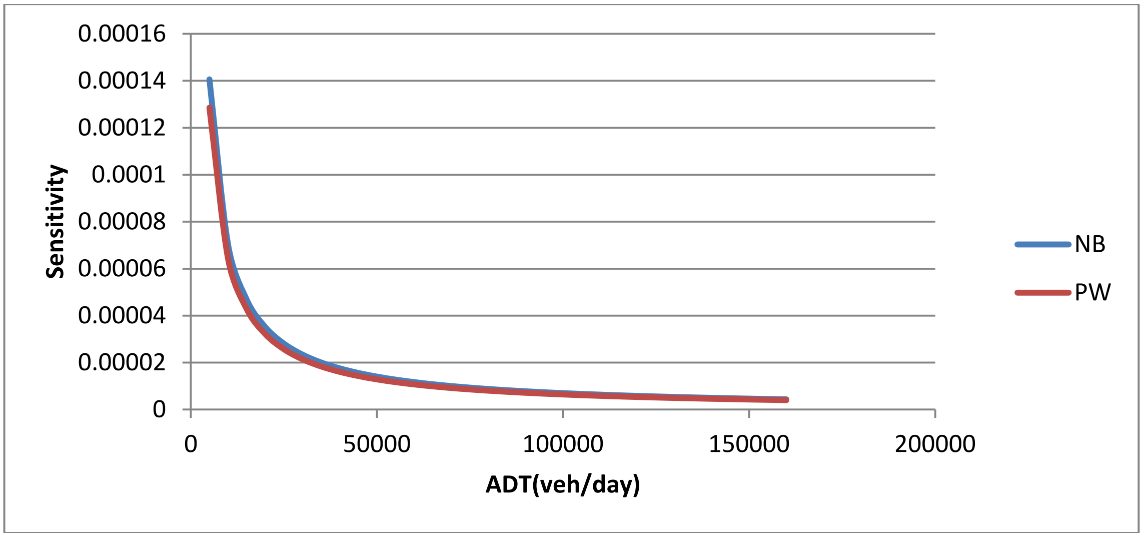


Fig. 5- 4. Sensitivity Analysis of Covariate ADT for Indiana Data

5.3.1.5 Model Validation

A cumulative residual (CURE) plot can be used to examine whether the model fits the data along the entire value range of one explanatory variable and also to ascertain whether a candidate explanatory variable should be included (Hauer and Bamfo 1997; Lin et al. 2002). A better fit occurs when the cumulative residuals oscillate around 0 against a certain variable. CURE plot can be also used to identify potential biases within the range of the variable investigated.

Fig. 5-5 illustrated the CURE plots for both the NB and PW model for Indiana data. It can be seen that the patterns for the NB and PW model are pretty similar. Both the models share similar flaws in predicting crash counts when variable ADT is larger than 30000 veh/day. Approximately from point A in the plot, both NB and PW lines are drifting down, and it shows that both models overestimate the crash counts against this range. Similar analysis can be conducted to other covariates and the figures are presented in Appendix D.

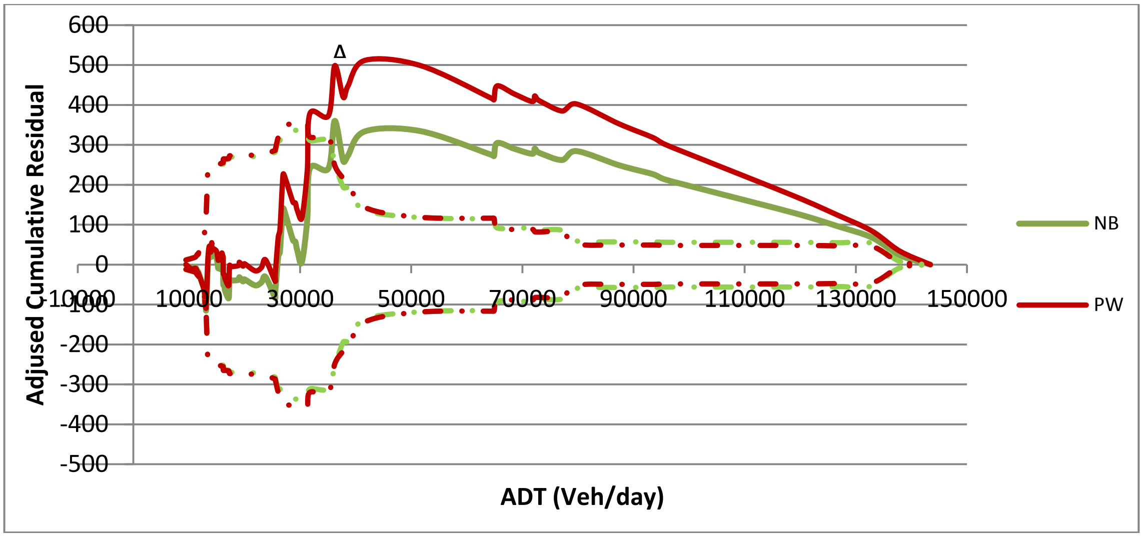


Fig.5-5. CURE Plot for Indiana Data (Dotted Lines represents ± 2 Std. Dev.)

5.3.1.6 Crash Variance

Based on the parameter estimates, the estimated crash variance and observed crash variance were calculated and plotted. The trend of crash variance against the mean is illustrated in Fig. 5-6. It shows that the model-based estimated variance for the PW model and NB model have a similar shape. And the estimated crash variance is very close to the observed variance especially when crash mean is below about 25 for both PW and NB models. When the data are characterized with higher mean, the PW model gives slightly higher predicted crash variance than the NB model. In general, the model-based estimated variance indicates that for both models, the estimated variance has a perfect monotone increase and it is highly correlated with the crash mean (Geedipally and Lord 2011). However, the observed crash variance for both PW and NB models is smaller than the predicted value when the crash mean is high.

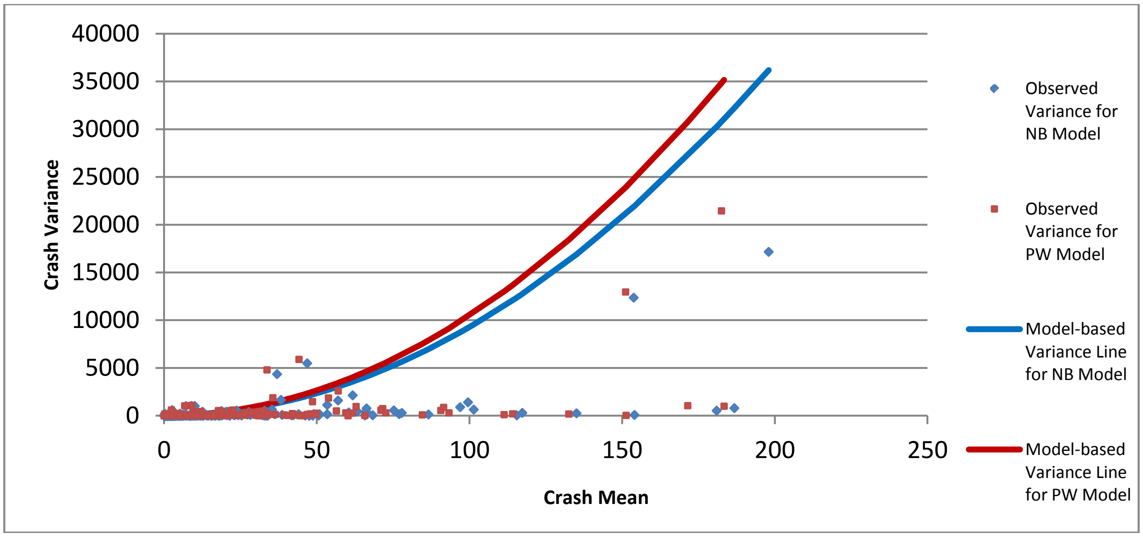


Fig. 5-6 Crash Variance vs. Crash Mean for Indiana Data

5.3.2 Texas Data

5.3.2.1 Parameter Estimate

Table 5-5 summarizes the estimated coefficients for Texas data:

Table 5- 5. Coefficient Estimates for the Texas Data

Variable	NB		PW	
	Value	S.E.	Value	S.E.
Intercept $(\ln \beta_0)$	-6.3590	0.3907	-6.2280	0.4298
ADT (β_1)	0.9770	0.0424	0.9682	0.0430
LW (β_2)	-0.0532	0.0169	-0.0509	0.0163
SW (β_3)	-0.0100	0.0033	-0.0101	0.0033
CD (β_4)	0.0674	0.0121	0.0694	0.0122
ϕ	2.5510	0.2338		
ω			1.0130	0.3068
k			1.6290	0.0817

As illustrated above, the functional form for the Texas data is similar to the one used for Indiana data. The segment length was considered as an offset in the model as well. All the coefficients are significant at 95% level. In general, all the estimated coefficients are logical and consistent with the existing literatures. The increase in the traffic flow and the horizontal curve density increases the crash risk, where the increase in lane width and shoulder width decreases the crash risk.

5.3.2.2 Goodness-of-fit Statistics

Table 5-6 presents the GOF statistics for the PW and NB models:

Table 5- 6. GOF Statistics for Texas Data

GOF	NB	PW
DIC	4784.45	4771.61
Pearson χ^2	1798.14	1869.32
MAD	1.70	1.69
MSPE	11.24	11.43

Based on the GOF statistics for Texas data, it can be seen that DIC value indicates that the PW model fits the data better than the NB model. However, the Pearson χ^2 value is in favor of the NB model. For the MAD and MSPE statistics, there are only slightly differences between those two models.

5.3.2.3 Estimation Accuracy

Fig. 5-7 and Fig.5-8 illustrate the comparison between observed and predicted data for two models. It is found the PW and NB model provide similar predictions with certain level of accuracy in terms of frequency distribution. For Texas data, the PW provides higher predicted values than the NB model, which is different from the case of the Indiana data. And when the observed crash counts are below 8, both PW and NB model are under-estimating the crash counts for most of time but the estimates are very close to the observed crash counts. When the observed crash counts keep increasing, both the PW and NB model are under-estimating the crash counts for most of the time.

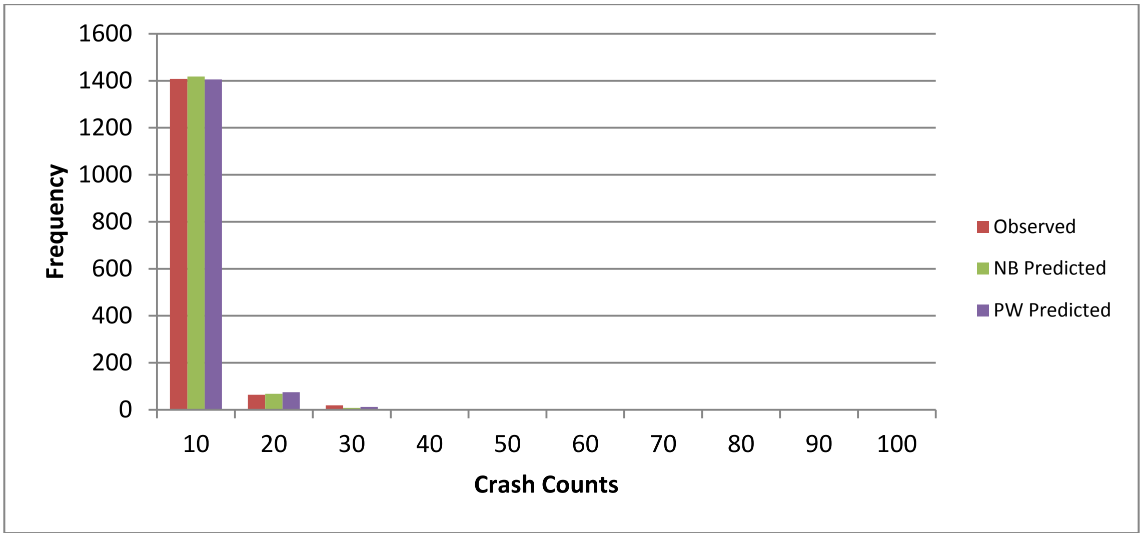


Fig. 5-7 Histograms of Observed and Predicted Crash Counts for Texas Data

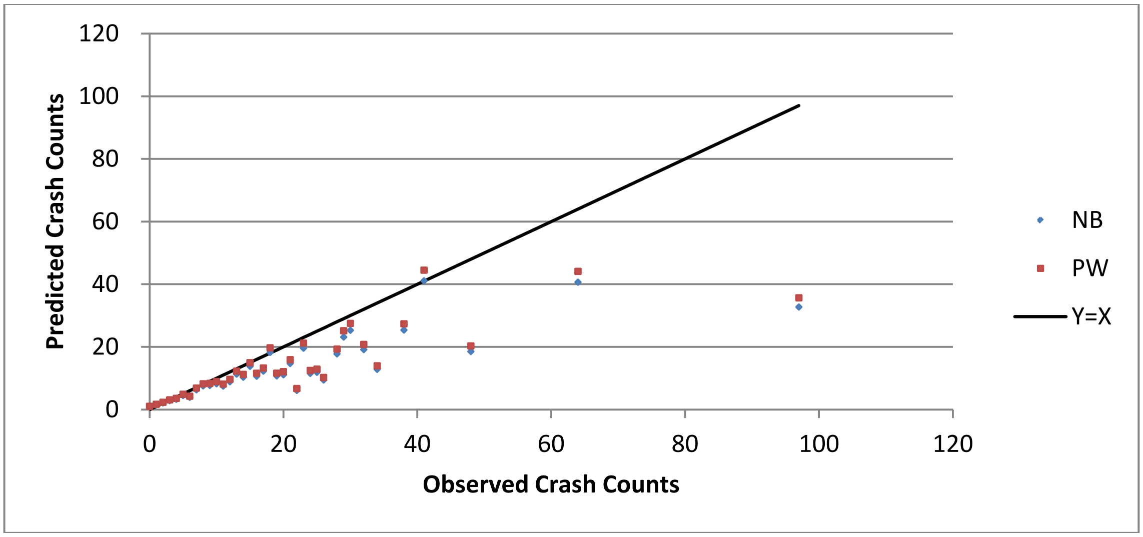


Fig. 5-8 Predicted Crash vs. Observed Crash for Texas Data

The average residual for the PW model is 0.18 and the minimum and maximum residuals are -13.19 and 65.32 respectively. For the NB model, the average, minimum and maximum residuals are 0.08, -13.70 and 64.26 respectively. The detailed table is in Appendix A.

Fig. 5-9 presents the predicted crash counts against the covariate ADT for two models. Unlike the Indiana data, in this case, the PW model provides the higher estimates than the NB model does. However, this difference is very small.

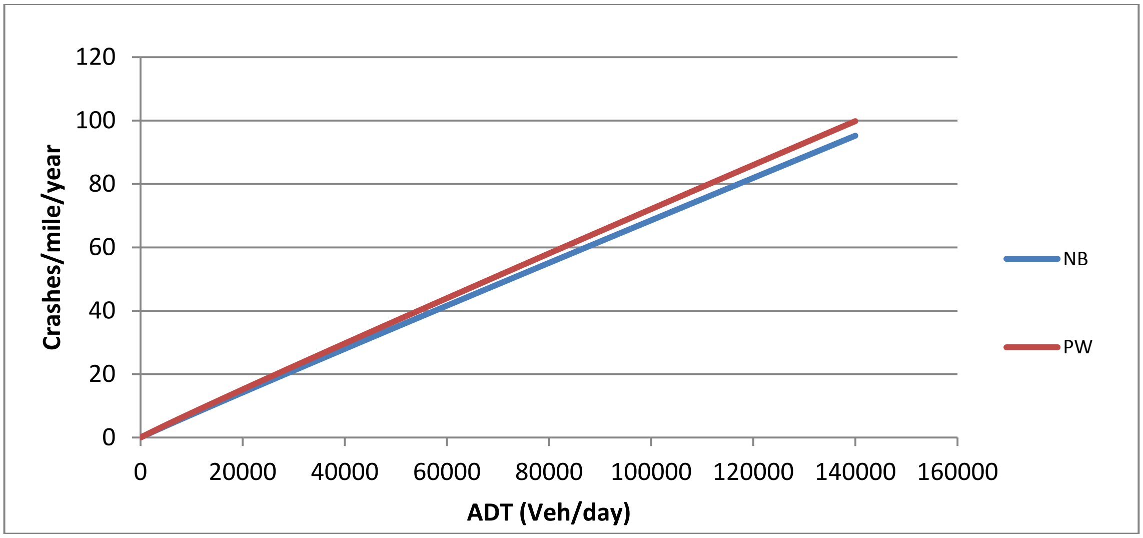


Fig. 5-9 Predicted Crash vs. ADT for Texas Data

5.3.2.4 Covariate Sensitivity

The sensitivity analysis for covariate ADT is illustrated in Fig. 5-10. For this dataset, the PW model and NB model have pretty similar performance.

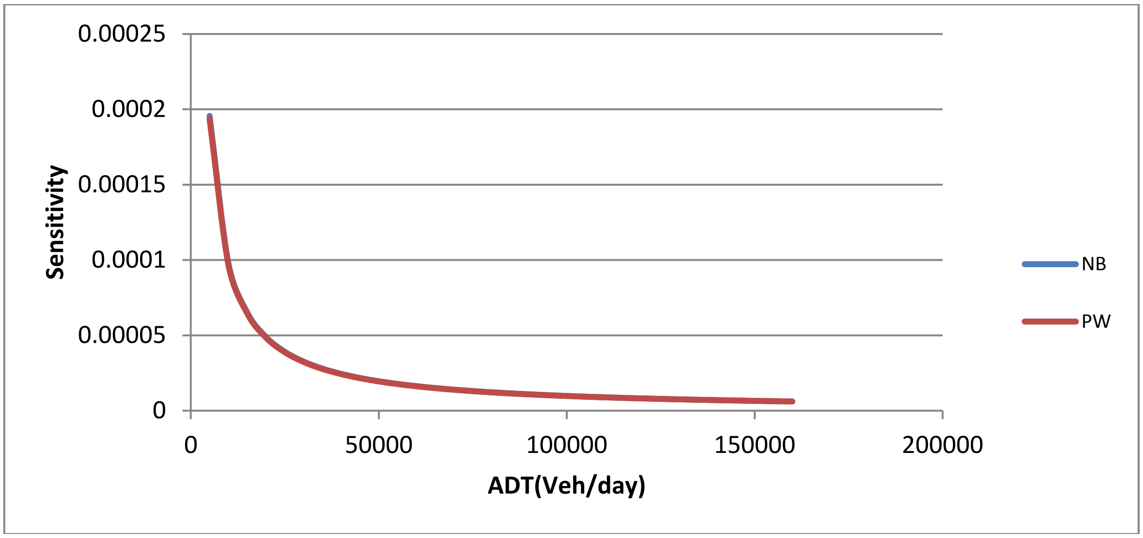


Fig. 5-10 Sensitivity Analysis of Covariate ADT for Texas Data

5.4.2.5 Model Validation

Fig. 5-11 illustrates the CURE plot for Texas data. It can be seen that for this dataset the PW model did not perform as well as the NB model. When ADT is between 5000 veh/day to 10000 veh/day, the PW model was overestimating the crash counts. Corresponding to the conclusion from Fig. 5-8, since the PW model gave higher estimates than the NB model, it is expected that the PW model oscillates farther than the NB model.

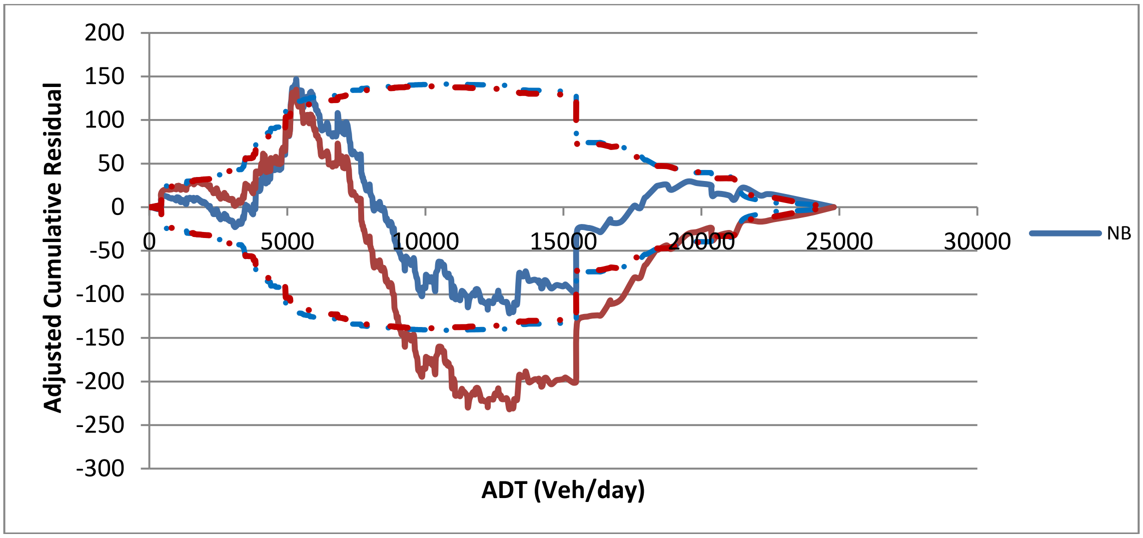


Fig. 5-11 CURE Plot for Texas Data (Dotted Lines represents ± 2 Std. Dev.)

5.4.2.6 Crash Variance

The estimated variance and observed variance are illustrated in Fig. 5-12. It indicates that the model-based estimated variances for the PW and NB model have almost identical shape. And for both PW and NB model, the crash variance has a monotone increase and it is correlated with the crash mean. As for the observed crash variance, it can be seen the observed crash variances for two models are similar. And when the crash mean is small, the estimated crash variance is close to the observed variance. Along with the increase of the crash mean, the observed crash variance for both PW and NB models is smaller than the estimated value.

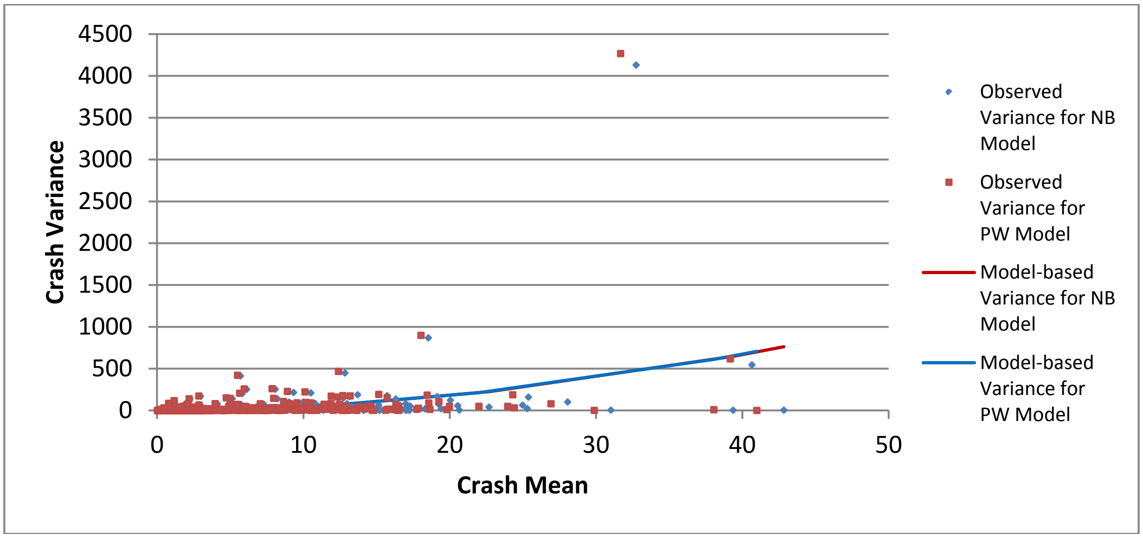


Fig. 5-12 Crash Variance vs. Crash Mean for Texas Data

5.4 Discussion

The simulation results have shown that the PW GLM has the potential for modeling vehicle crashes when compared to the NB GLM. This section has applied the PW model to the observed data. The following discussions can be observed based on the results above:

1) First, the PW model performs as well as the NB model for the functional form that includes several covariates based on the parameter estimates for two

datasets. Both PW and NB model provide similar estimates and those values are consistent with existing literature.

2) Second, it is found that for both datasets the GOF statistics have given different preferences towards these two models. The DIC values for both datasets indicate that the PW GLM provided better fit to the observed data. However, MSPE statistic prefers the NB GLM. The Pearson χ^2 and MAD statistics shows different preference in those two cases. As Miaou and Lord stated in their research, the model performance cannot be only judged by various GOF statistics, “goodness-of-logic” also needs to be considered (Miaou and Lord 2003). So we cannot solely reply on one or two GOF statistics to conclude that which model is better.

3) Based on the comparisons from prediction accuracy, covariate sensitivity analyzes and relation between crash variance and mean, the PW GLM has the similar performance as the NB model. However, based on the CURE plots against covariate ADT, it is found the PW model didn’t perform as well as the NB model, especially for the Texas dataset. It might share the model flaws with the NB model in overestimating the crash counts within certain flow range.

4) Even though the computational time was not documented in the results above, the time for both GLMSs were quite similar in WinBUGS with only 10 to 20 seconds difference. The computational time for estimating the parameters varies due to the different computer facilities. One thing that should be noted is the

PW GLM has one extra parameter and 10 to 20 seconds of extra computational time for the PW GLM is reasonable and acceptable.

5) Due to the quality and availability of observed data, the datasets used in this study do not fully cover the equi-disperison and over-dispersion situations. As NB GLMs have difficulties in handling highly dispersed data, the modeling performance of PW GLMs needs to be further examined under those situations.

5.5 Summary

This section has documented the modeling performance of the PW GLM on two observed datasets. The modeling performance of the PW GLM was assessed through a series of comparison analyses with the NB model. The parameter estimates for the two GLMs were very close and both models had similar performances in sensitivity, model-checking and mean-variance analyses. Therefore, the comparison results showed that the PW GLM performed as well as the NB GLM for the given datasets. The next section presents the conclusions and future work related to the PW model.

6. CONCLUSIONS AND FUTURE WORK

This section summarizes the work accomplished in this research and conclusions that are obtained from the study. Recommendations for the future work are also provided.

6.1 Conclusions

The primary objective of this research is to develop and evaluate the modeling performance of the PW GLM. PW model is an innovative model that has never been used in model traffic crash counts. To examine the PW GLM modeling performance under different dispersion level, the following tasks were completed in this research:

- Since PW model is an innovative statistical model, its statistical characteristics have been defined in this study. PW model is assumed to be a mixture of Poisson and Weibull distributions. Therefore, its p.d.f , mean and variance were developed based on this assumption. However, the p.d.f of the PW model has no closed form, and thus to estimate the parameters, Bayesian estimation approach was used.
- To initially examine the modeling performance of the PW GLM, a series of simulated dataset was generated corresponding to different dispersion levels. The evaluation was conducted by comparing the true parameters with the estimated parameters. It was found that the PW GLM was able to reproduce the true parameter values with considerably high accuracy.

- After initial assessment on simulated data, PW GLMs was applied to the observed data. Two datasets were used and the functional form including several covariates was selected to model the crash counts. The modeling performance of the PW GLMs was compared with the NB GLMs. By the GOF statistics, it was found that the PW model performed as well as the NB model.

6.2 Future Work

Given the fact that the PW GLM was first introduced in modeling crash data in transportation area, there are many lines of research activities that could be investigated in the future work.

First, in this study, for both simulated and observed datasets, the sample size and sample mean were given the moderate levels. It is known that crash data sometime can be subjected to small-sample-size and low-sample-mean issues which the NB model cannot handle very well. For such datasets, the inverse dispersion parameter can be significantly biased or mis-estimated (Lord 2006). This can negatively influence the standard errors associated with the model's coefficients. For the empirical Bayes (EB) estimate method, the outcome might also be problematic. Further work about the PW model therefore is needed on stability when data are characterized by small-sample-size and low-sample-mean values.

Furthermore, as EB method is now frequently used in highway safety analyzes, an EB modeling framework also should be developed. The PW GLM might also have the potential to identify hazardous sites. Therefore, the framework of how to apply PW GLM is also needed to be further investigated.

For this study, it is assumed that all the covariates are independent. Further study could be done to examine the effects of covariate-dependent parameters.

Last, a well-defined likelihood function and the related moments for the PW model should be developed, if it is possible to developed one. This way, the maximum likelihood estimation (MLE) method could be used for estimating PW GLMs.

REFERENCES

Anastasopoulos, P. C., Tarko, A. P., and Mannering, F. L. (2008). "Tobit analysis of vehicle accident rates on interstate highways." *Accid. Anal. Prev.,* 40(2), 768-775.

Aptel, I., Salmi, L. R., Masson, F., Bourdé, A., Henrion, G., and Erny, P. (1999). "Road accident statistics: discrepancies between police and hospital data in a French island." *Accid. Anal. Prev.,* 31(1-2), 101-108.

Carson, J., and Mannering, F. (2001). "The effect of ice warning signs on ice-accident frequencies and severities." *Accid. Anal. Prev.,* 33(1), 99-109.

Casella, G., and Berger, R. L. (2001). *Statistical Inference, 2nd Ed.*, Duxbury Press, Belmont, CA.

Cohen, A. (1963). "Estimation in mixtures of discrete distributions." *Proc., Int. Symp. on Classical and Contagious Discrete Distributions,* Pergamon Press, New York,351-372.

Conway, R. W., and Maxwell, W. L. (1962). "A queuing model with state dependent service rates." *J. Ind. Eng.,* 12(2), 132-136.

Elvik, R., and Mysen, A. B. (1999). "Incomplete accident reporting: meta-analysis of studies made in 13 countries." *Transp. Res. Rec.,* 1665, 133-140.

Geedipally, S. R. (2009). "Examining the application of conway-maxwell-poisson models for analyzing traffic crash data." doctoral dissertation, Texas A&M University, College Station, TX.

Geedipally, S. R., and Lord, D. (2010). "Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson-gamma models." *Accid. Anal. Prev.*, 42(4), 1273-1282.

Geedipally, S. R., and Lord, D. (2011). "Examining the Crash Variances Estimated by the Poisson-Gamma and Conway-Maxwell-Poisson Models." *Proc., 90th Annual Meeting of the Transportation Research Board,* Washington, DC.

Geedipally, S. R., Lord, D., and Dhavala, S. S. (2011). "The negative binomial-Lindley generalized linear model: characteristics and application using crash data." *Accid. Anal. Prev.*, 45,258-265.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice,* Chapman & Hall/CRC, Boca Raton, Florida.

Hauer, E. (1997). *Observational Before--After Studies in Road Safety*, Pergamon, Elsevier Science, Oxford, UK.

Hauer, E., and Bamfo, J. (1997). "Two tools for finding what function links the dependent variable to the explanatory variables." *Proc., ICTCT 1997 Conference,* Lund, Sweden.

Hauer, E., and Hakkert, A. (1988). "Extent and some implications of incomplete accident reporting." *Transp. Res. Rec.*, (1185), 1-11.

Kadane, J. B., Shmueli, G., Minka, T. P., Borle, S., and Boatwright, P. (2006). "Conjugate analysis of the Conway-Maxwell-Poisson distribution." *Bayesian Anal.* 1(2), 363-374.

Kumara, S., and Chin, H. (2003). "Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros." *Traffic Inj. Prev.*, 4(1), 53-57.

Lambert, D. (1992). "Zero-inflated Poisson regression, with an application to defects in manufacturing." *Technometrics*, 34, 1-14.

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

Lin, D., Wei, L., and Ying, Z. (2002). "Model-checking techniques based on cumulative residuals." *Biometrics*, 58(1), 1-12.

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." *Accid. Anal. Prev.*, 38(4), 751-766.

Lord, D., and Bonneson, J. A. (2005). "Calibration of predictive models for estimating safety of ramp design configurations." *Transp. Res. Rec.,* 1908, 88-95.

Lord, D., and Mahlawat, M. (2009). "Examining Application of Aggregated and Disaggregated Poisson-Gamma Models Subjected to Low Sample Mean Bias." *Transp. Res. Rec.*, 2136, 1-10.

Lord, D., and Mannering, F. (2010). "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives." *Transp. Res. Part A: Policy and Practice,* 44(5), 291-305.

Lord, D., and Persaud, B. N. (2000). "Accident prediction models with and without trend: application of the generalized estimating equations procedure." *Transp. Res. Rec.,* 1717, 102-108.

Lord, D., Guikema, S. D., and Geedipally, S. R. (2008b). "Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes." *Accid. Anal. Prev.,* 40(3), 1123-1134.

Lord, D., Washington, S. P., and Ivan, J. N. (2005). "Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory." *Accid. Anal. Prev.*, 37(1), 35-46.

Lord, D., Geedipally, S., Persaud, B., Washington, S., van Schalkwyk, I., Ivan, J., Lyon, C., and Jonsson, T. (2008a). *NCHRP Web-Only Document 126: Methodology to*

Predict the Safety Performance of Rural Multilane Highways. Transportation Research Board of the National Academies, Washington, D.C..

Maher, M. J., and Summersgill, I. (1996). "A comprehensive methodology for the fitting of predictive accident models." *Accid. Anal. Prev.*, 28(3), 281-296.

Maycock, G., and Hall, R.D. (1984). *Accidents at 4-arm Roundabouts*. TRRL Laboratory Report 1120, Transportation and Road Research Laboratory, Crowthorne, Berkshire, UK.

Miaou, S. P., and Lord, D. (2003). "Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods." *Transp. Res. Rec.,*1840, 31-40.

Miaou, S. P., and Song, J. J. (2005). "Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence." *Accid. Anal. Prev.*, 37(4), 699-720.

Mitra, S., and Washington, S. (2007). "On the nature of over-dispersion in motor vehicle crash prediction models." *Accid. Anal. Prev.*, 39(3), 459-468.

Montgomery, D. C., and Runger, G. C. (2003). *Applied Statistics and Probability for Engineers*. John Wiely & Sons Inc, Hoboken, NJ.

Oh, J., Washington, S. P., and Nam, D. (2006). "Accident prediction model for railway-highway interfaces." *Accid. Anal. Prev.*, 38(2), 346-356.

Park, E. S., and Lord, D. (2007). "Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity." *Transp. Res. Rec.,* 2019, 1-6.

Rider, P. R. (1961). "Estimating the parameters of mixed Poisson, binomial and Weibull distributions by the method of moments." *Bulletin De l'Institut International De Statistiques*, 38, Part 2.

Shankar, V., Milton, J., and Mannering, F. (1997). "Modeling accident frequencies as zero-altered probability processes: An empirical inquiry." *Accid. Anal. Prev.*, 29(6), 829-837.

Shankar, V. N., Albin, R. B., Milton, J. C., and Mannering, F. L. (1998). "Evaluating median crossover likelihoods with clustered accident counts: An empirical inquiry using the random effects negative binomial model." *Transp. Res. Rec.*, 1635, 44-48.

Shankar, V. N., Ulfarsson, G. F., Pendyala, R. M., and Nebergall, M. L. B. (2003). "Modeling crashes involving pedestrians and motorized traffic." *Saf. Sci.*, 41(7), 627-640.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). "A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution." *J. R. Stat. Soc.: Series C (Applied Statistics)*, 54(1), 127-142.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *J. R. Stat. Soc.: Series B (Statistical Methodology),* 64(4), 583-639.

Washington, S. P., Karlaftis, M. G., and Mannering, F. L. (2009). *Statistical and Economic Methods for Transportation Data Analysis." 2nd Ed*.. Chapman & Hall/CRC, Boca Raton, FL.

Wood, G. (2002). "Generalised linear accident models and goodness of fit testing." *Accid. Anal. Prev.*, 34(4), 417-427.

APPENDIX A

COMPARISON BETWEEN PREDICTED AND OBSERVED CRASH COUNTS

- **Indiana Data**

Observed Crash Count	NB				PW			
	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual
0	2.004	-2.004	-13.069	-0.026	1.876	-1.876	-12.717	-0.025
1	2.938	-1.938	-7.775	0.233	2.813	-1.813	-7.390	0.257
2	4.025	-2.025	-7.605	1.526	3.787	-1.787	-6.631	1.565
3	3.737	-0.737	-7.257	2.253	3.592	-0.592	-6.994	2.315
4	5.317	-1.317	-5.852	2.377	5.165	-1.165	-5.472	2.414
5	7.172	-2.172	-8.933	0.731	6.993	-1.993	-8.807	0.849
6	8.282	-2.282	-13.961	3.517	7.852	-1.852	-12.407	3.735
7	11.155	-4.155	-15.527	4.817	10.762	-3.762	-14.122	4.866
8	9.672	-1.672	-19.744	5.648	9.236	-1.236	-18.583	5.848
9	10.696	-1.696	-9.994	5.807	10.214	-1.214	-9.237	6.013
10	23.339	-13.339	-25.344	0.848	21.843	-11.843	-22.314	1.187
11	7.681	3.319	-6.552	10.321	7.282	3.718	-6.264	10.350
12	9.993	2.007	1.922	2.092	9.437	2.563	2.271	2.854
13	14.842	-1.842	-10.465	9.651	14.498	-1.498	-10.113	9.822
14	16.792	-2.792	-13.832	13.751	16.126	-2.126	-12.607	13.751
15	15.421	-0.421	-12.384	-12.384	14.794	0.206	-12.134	-12.134
16	6.391	9.609	9.609	9.609	6.341	9.659	9.659	9.659
17	12.074	4.926	1.493	10.743	11.679	5.321	1.353	10.828
18	18.173	-0.173	-0.173	-0.173	17.216	0.784	0.784	0.784
19	21.118	-2.118	-9.243	5.007	19.918	-0.918	-7.609	5.773
20	5.165	14.835	14.835	14.835	4.665	15.335	15.335	15.335
21	12.434	8.566	8.566	8.566	12.212	8.788	8.788	8.788
22	16.507	5.493	0.495	11.793	15.316	6.684	1.646	12.614
24	29.106	-5.106	-12.291	5.127	27.290	-3.290	-10.018	6.361
27	6.533	20.467	16.562	6.533	5.939	21.061	17.529	5.939
30	22.992	7.008	7.008	7.008	21.591	8.409	8.409	8.409

Observed Crash Count	NB				PW			
	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual
31	25.389	5.611	5.611	5.611	23.864	7.136	7.136	7.136
32	23.391	8.609	-1.235	19.377	21.443	10.557	1.017	20.902
33	12.592	20.408	20.408	20.408	11.988	21.012	21.012	21.012
35	38.003	-3.003	-7.105	38.003	35.816	-0.816	-5.482	35.816
36	23.155	12.845	12.845	12.845	22.550	13.450	13.450	13.450
38	50.872	-12.872	-12.872	-12.872	49.763	-11.763	-11.763	-11.763
39	22.365	16.635	6.306	31.854	21.847	17.153	6.165	32.229
41	16.046	24.954	20.584	32.070	15.017	25.983	22.165	32.562
42	36.672	5.328	-21.289	31.944	34.276	7.724	-17.503	32.950
45	41.751	3.249	-21.146	21.743	39.362	5.638	-16.704	23.199
47	35.681	11.319	-0.564	23.201	33.514	13.486	2.537	24.434
48	34.615	13.385	9.342	17.427	32.527	15.473	11.600	19.346
49	41.754	7.246	7.246	7.246	39.363	9.637	9.637	9.637
51	46.234	4.766	4.766	4.766	44.539	6.461	4.766	4.766
52	38.454	13.546	9.519	17.574	37.992	14.008	9.514	18.503
57	48.431	8.569	6.384	12.938	45.190	11.810	10.218	14.993
60	65.787	-5.787	-5.787	-5.787	60.256	-0.256	-0.256	-0.256
61	68.319	-7.319	-7.319	-7.319	65.769	-4.769	-4.769	-4.769
62	74.285	-12.285	-37.530	12.960	69.705	-7.705	-29.555	14.146
66	53.430	12.570	12.570	12.570	49.700	16.300	16.300	16.300
67	96.960	-29.960	-29.960	-29.960	90.629	-23.629	-23.629	-23.629
75	86.585	-11.585	-11.585	-11.585	84.578	-9.578	-9.578	-9.578
76	101.408	-25.408	-25.408	-25.408	93.296	-17.296	-17.296	-17.296
79	49.375	29.625	18.521	40.729	46.021	32.979	22.576	43.382
87	53.430	33.570	33.570	33.570	48.599	38.401	38.401	48.599
90	76.986	13.014	13.014	13.014	72.622	17.378	17.378	72.622
94	66.306	27.694	27.694	27.694	62.881	31.119	31.119	62.881
95	77.858	17.142	17.142	17.142	71.017	23.983	23.983	71.017
97	56.973	40.027	40.027	40.027	53.839	43.161	43.161	53.839
99	75.237	23.763	23.763	23.763	71.595	27.405	27.405	71.595
100	117.311	-17.311	-17.311	-17.311	114.152	-14.152	-14.152	114.152
103	36.996	66.004	66.004	66.004	33.701	69.299	69.299	33.701
108	61.757	46.243	46.243	46.243	57.027	50.973	50.973	57.027
119	135.063	-16.063	-16.063	-16.063	132.534	-13.534	-13.534	132.534
121	46.865	74.135	74.135	74.135	44.175	76.825	76.825	44.175

Observed Crash Count	NB				PW			
	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual
122	115.511	6.489	6.489	6.489	111.343	10.657	10.657	111.343
145	154.048	-9.048	-9.048	-9.048	151.304	-6.304	-6.304	151.304
204	180.954	23.046	23.046	23.046	171.493	32.507	32.507	171.493
215	186.752	28.248	28.248	28.248	183.389	31.611	31.611	183.389
265	153.785	111.215	111.215	111.215	151.146	113.854	113.854	151.146
329	197.986	131.014	131.014	131.014	182.519	146.481	146.481	182.519

- **Texas Data**

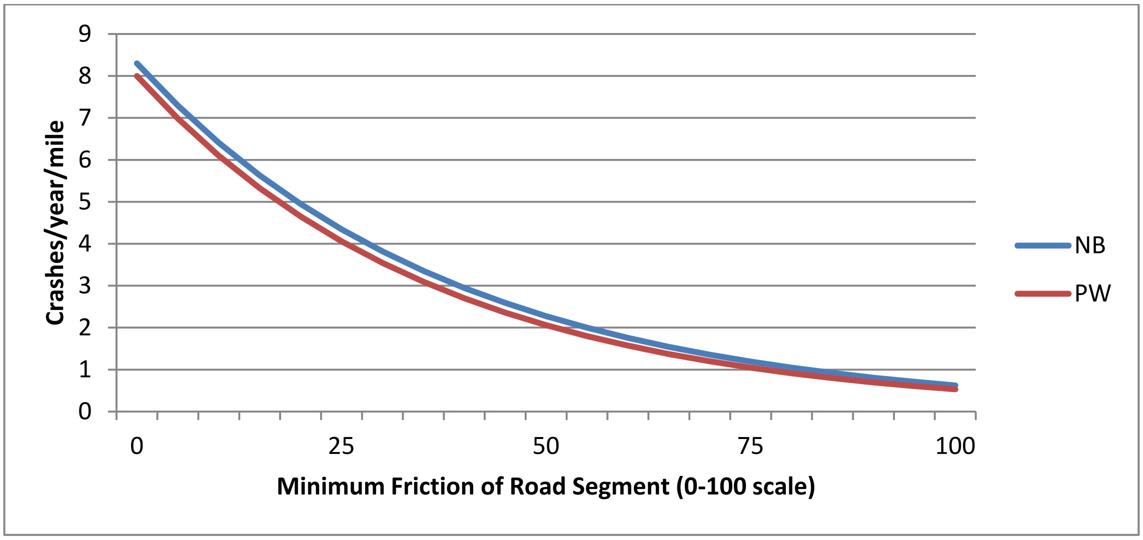
Observed Crash Count	NB				PW			
	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual
0	0.952	-0.952	-13.708	-0.018	1.039	-1.039	-14.857	-0.020
1	1.547	-0.547	-8.038	0.914	1.686	-0.686	-8.800	0.906
2	2.127	-0.127	-8.729	1.756	2.315	-0.315	-9.673	1.727
3	2.842	0.158	-6.048	2.809	3.092	-0.092	-6.816	2.790
4	3.285	0.715	-8.983	3.709	3.574	0.426	-10.047	3.672
5	4.488	0.512	-7.332	4.235	4.879	0.121	-8.383	4.165
6	3.913	2.087	-5.609	5.524	4.260	1.740	-6.534	5.477
7	6.306	0.694	-8.137	6.075	6.856	0.144	-9.354	5.988
8	7.553	0.447	-8.995	6.475	8.205	-0.205	-10.375	6.350
9	7.703	1.297	-11.044	7.089	8.362	0.638	-12.685	6.939
10	8.237	1.763	-7.286	9.181	8.947	1.053	-8.595	9.115
11	7.488	3.512	-1.561	8.836	8.136	2.864	-2.570	8.633
12	8.875	3.125	-4.445	10.800	9.638	2.362	-5.871	10.684
13	11.314	1.686	-7.547	8.878	12.265	0.735	-9.501	8.532
14	10.327	3.673	-4.311	11.754	11.223	2.777	-5.976	11.532
15	13.801	1.199	-4.420	4.766	14.954	0.046	-5.993	3.866
16	10.638	5.362	-1.021	13.034	11.587	4.413	-2.561	12.785
17	12.269	4.731	-7.977	12.095	13.267	3.733	-9.984	11.674

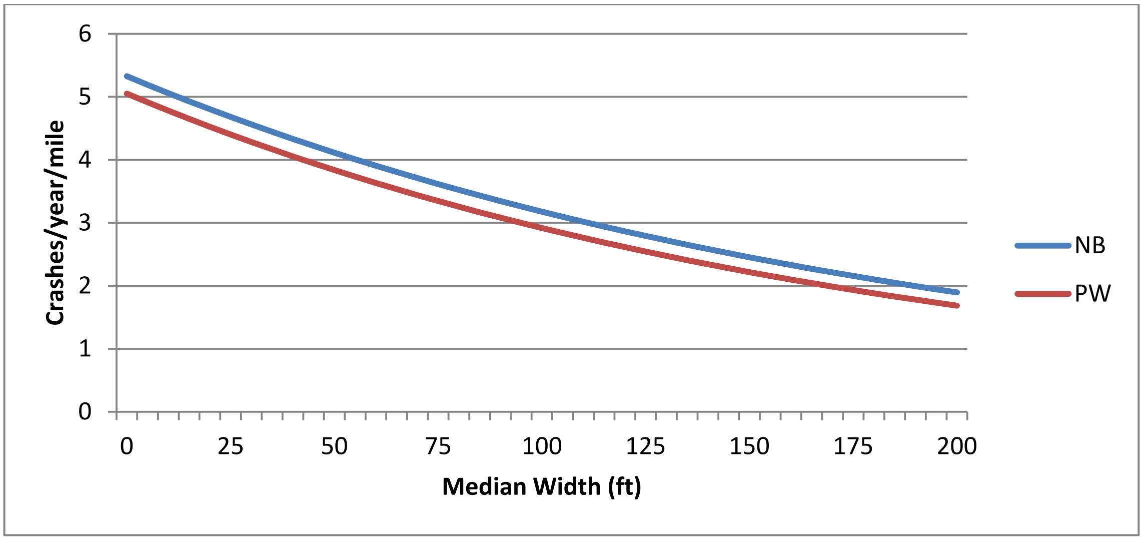
Observed Crash Count	NB				PW			
	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual	Ave. Predicted	Ave. Residual	Min. Residual	Max. Residual
18	18.226	-0.226	-10.061	8.628	19.702	-1.702	-12.312	7.802
19	10.733	8.267	6.554	9.979	11.602	7.398	5.524	9.273
20	11.103	8.897	2.739	14.183	12.091	7.909	1.355	13.643
21	14.711	6.289	3.997	8.580	15.911	5.089	2.556	7.623
22	6.120	6.120	6.120	6.120	6.706	6.706	6.706	6.706
23	19.593	3.407	2.345	4.468	21.206	1.794	0.705	2.884
24	11.542	10.726	6.779	15.922	12.478	9.710	5.448	15.144
25	11.886	13.114	12.264	14.492	12.895	12.105	11.096	13.619
26	9.431	16.569	12.835	20.303	10.240	15.760	11.712	19.809
28	17.776	10.224	8.762	11.686	19.307	8.693	7.091	10.294
29	23.139	5.861	-2.013	13.281	25.154	3.846	-4.640	11.943
30	25.299	4.701	4.701	25.299	27.489	2.511	2.511	27.489
32	19.143	12.857	12.857	19.143	20.782	11.218	11.218	20.782
34	12.841	21.159	21.159	12.841	13.959	20.041	20.041	13.959
38	25.379	12.621	12.621	25.379	27.367	10.633	10.633	27.367
41	41.103	-0.103	-1.837	1.630	44.497	-3.497	-5.151	-1.843
48	18.539	29.461	29.461	29.461	20.308	27.692	27.692	27.692
64	40.652	23.348	23.348	23.348	44.105	19.895	19.895	19.895
97	32.737	64.263	64.263	64.263	35.668	61.332	61.332	61.332

APPENDIX B

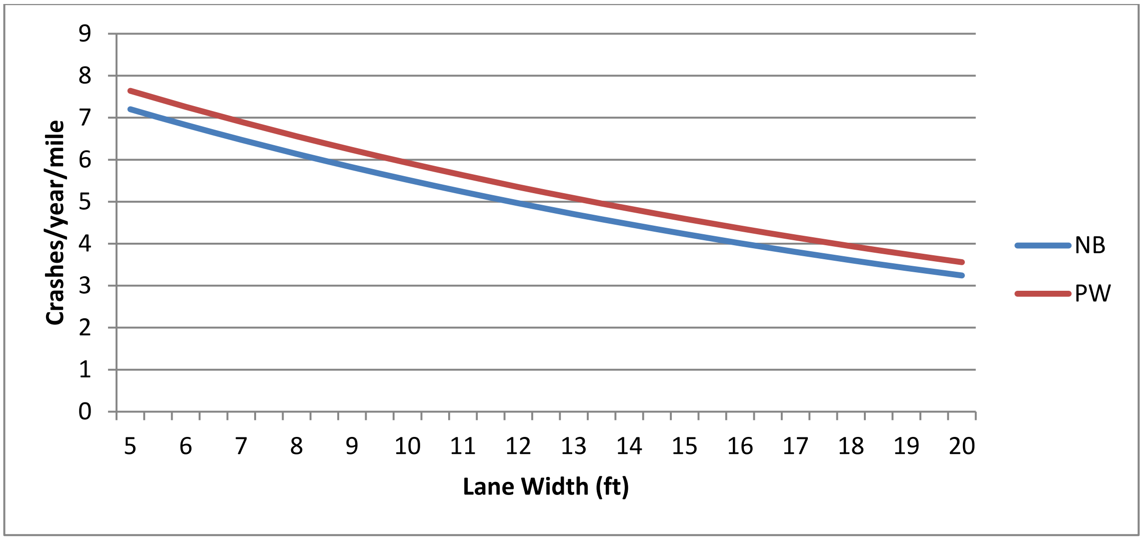
PREDICTED CRASH COUNTS AGAINST OTHER COVARIATES

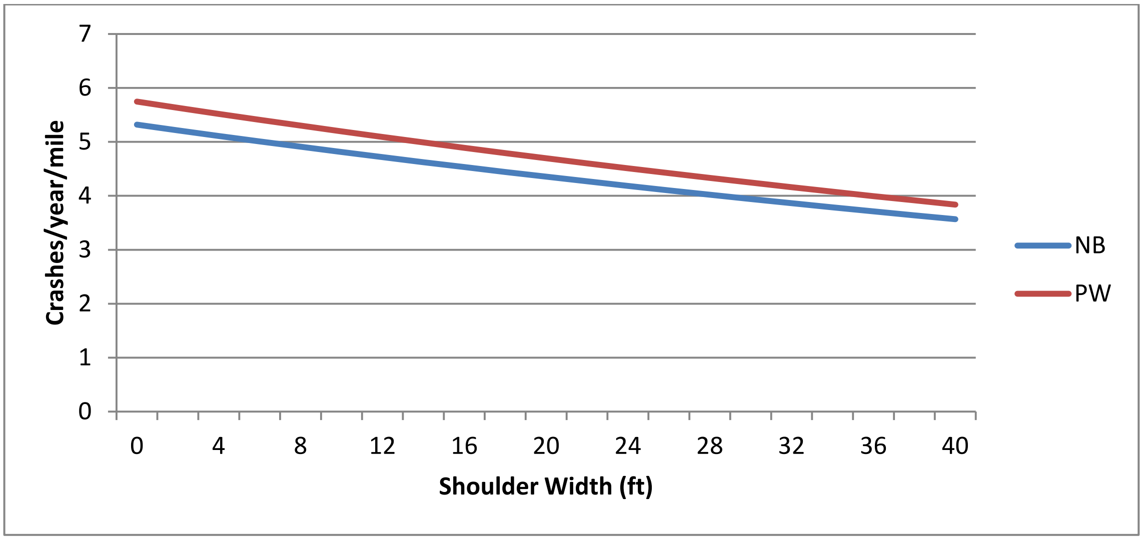
- **Indiana Data**

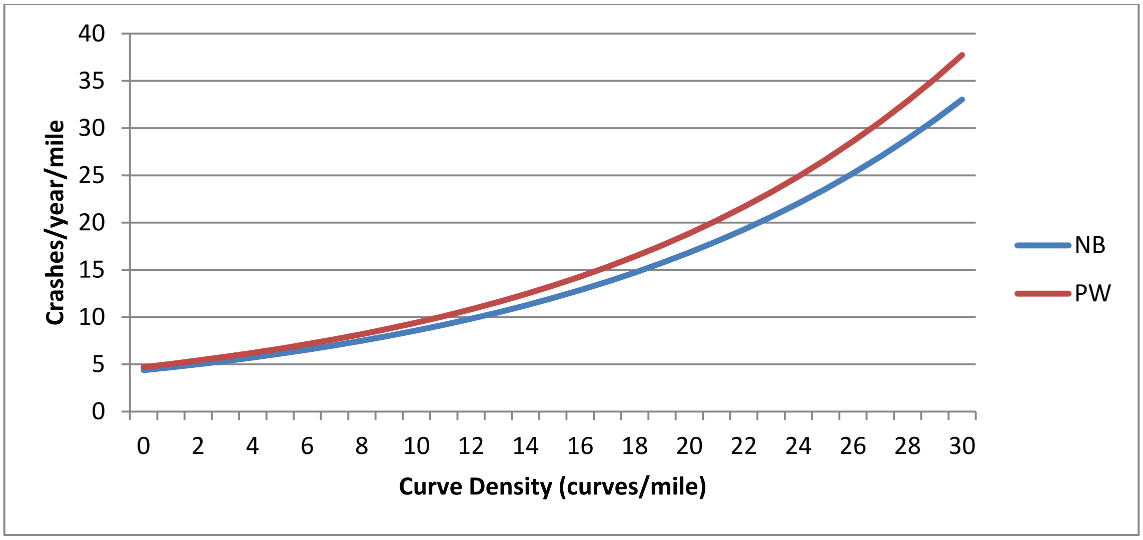




- **Texas Data**


Crashes/year/mile
0
1
2
3
4
5
6
7
8
9
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Lane Width (ft)
NB
PW

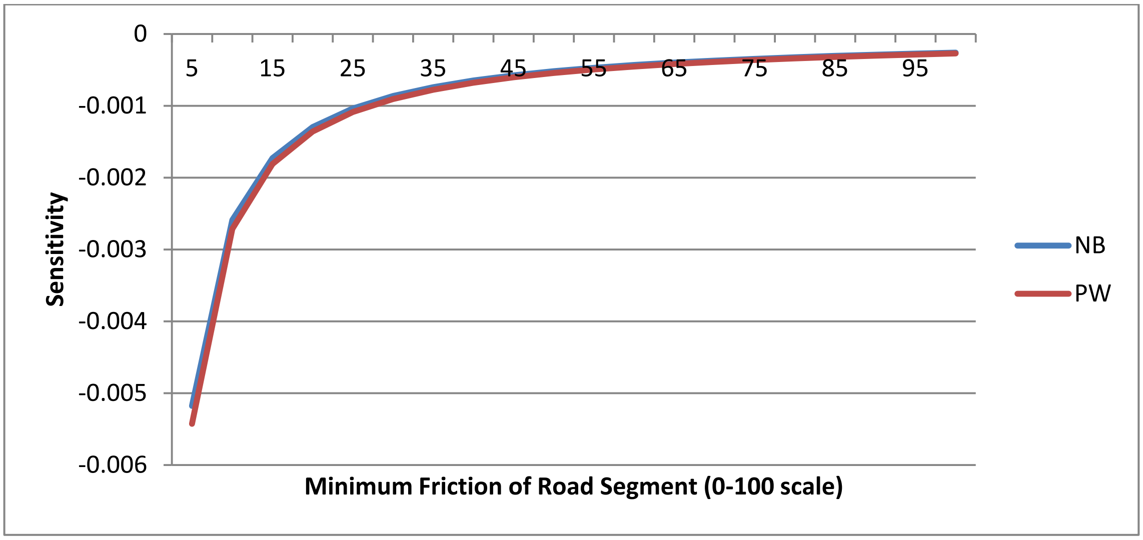

Crashes/year/mile
0
1
2
3
4
5
6
7
0
4
8
12
16
20
24
28
32
36
40
Shoulder Width (ft)
NB
PW

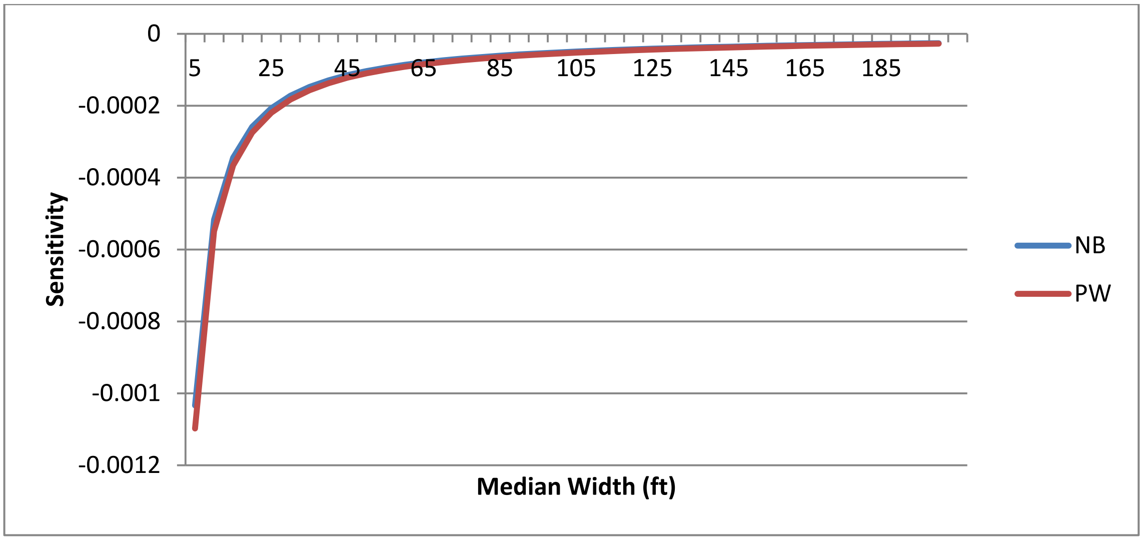

Crashes/year/mile
40
35
30
25
20
15
10
5
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Curve Density (curves/mile)
NB
PW

APPENDIX C

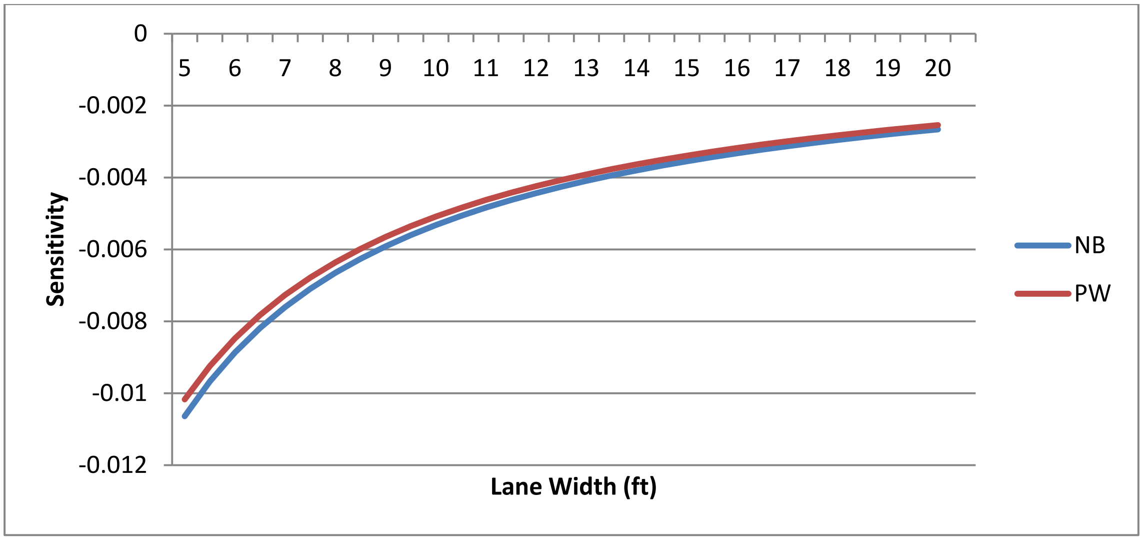
SENSITIVITY ANALYSIS OF OTHER COVARIATES

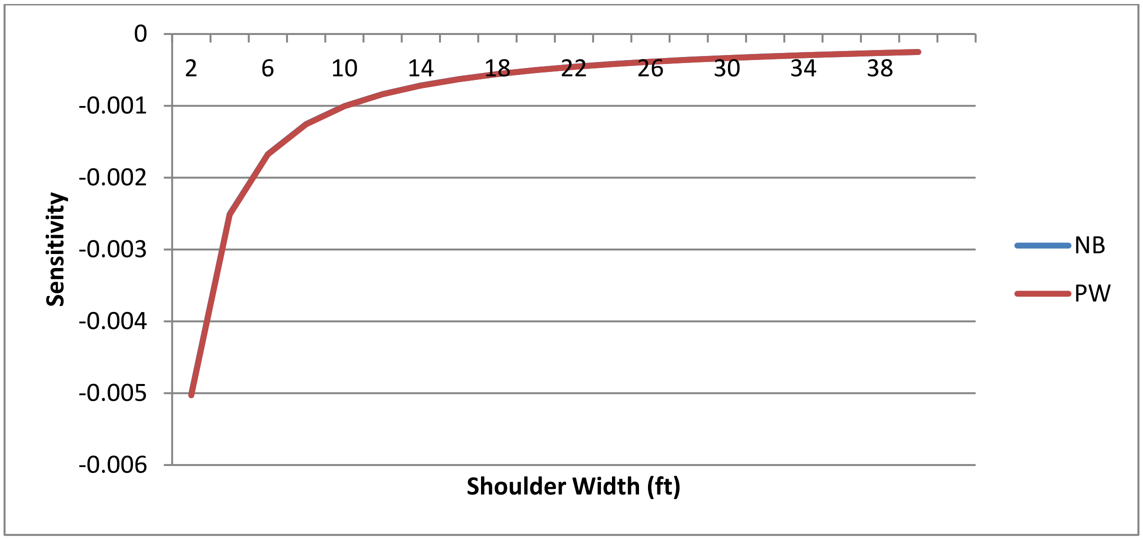
- **Indiana Data**

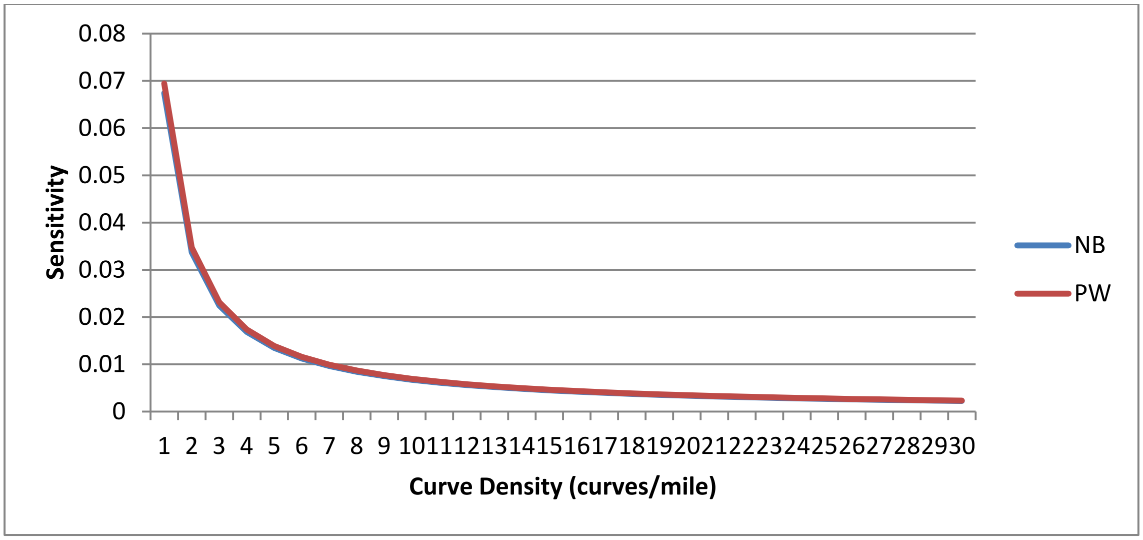




- **Texas Data**


0
-0.002
-0.004
-0.006
-0.008
-0.01
-0.012
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Sensitivity
Lane Width (ft)
NB
PW

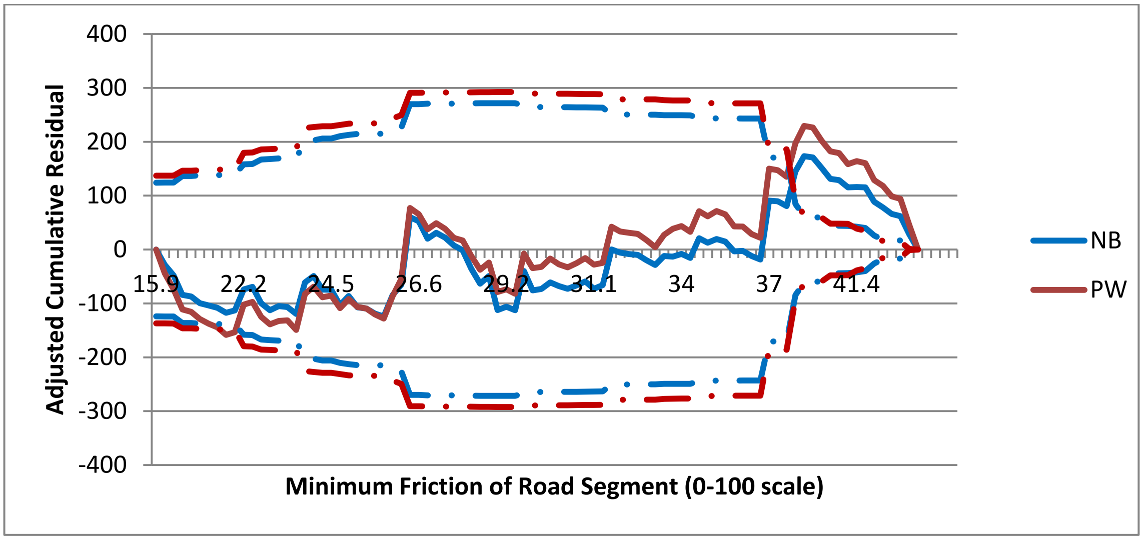

0
-0.001
-0.002
-0.003
-0.004
-0.005
-0.006
2
6
10
14
18
22
26
30
34
38
Sensitivity
Shoulder Width (ft)
NB
PW

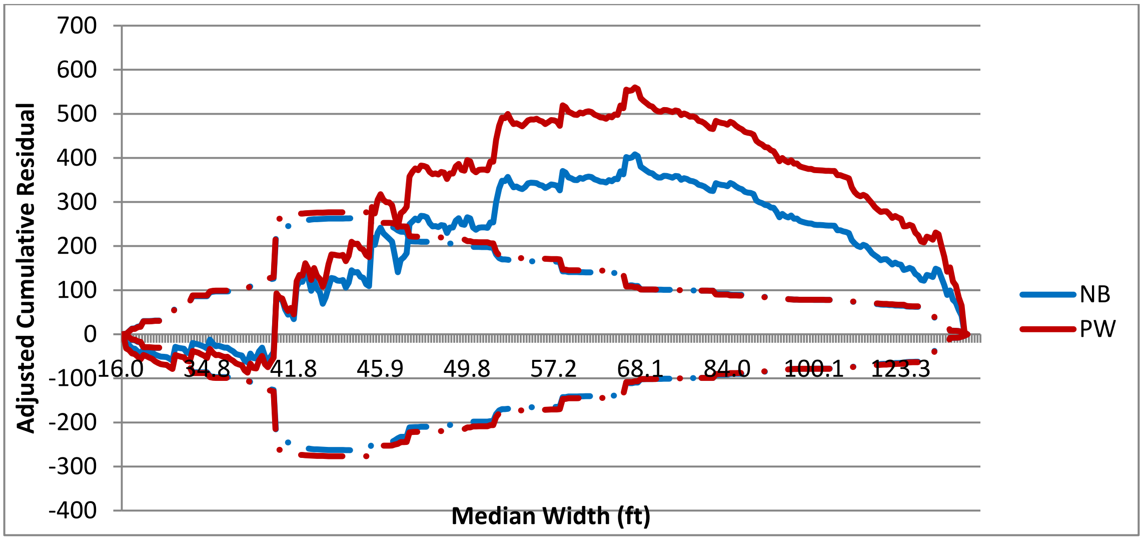

0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Sensitivity
1 2 3 4 5 6 7 8 9 101112131415161718192021222324252627282930
Curve Density (curves/mile)
NB
PW

APPENDIX D

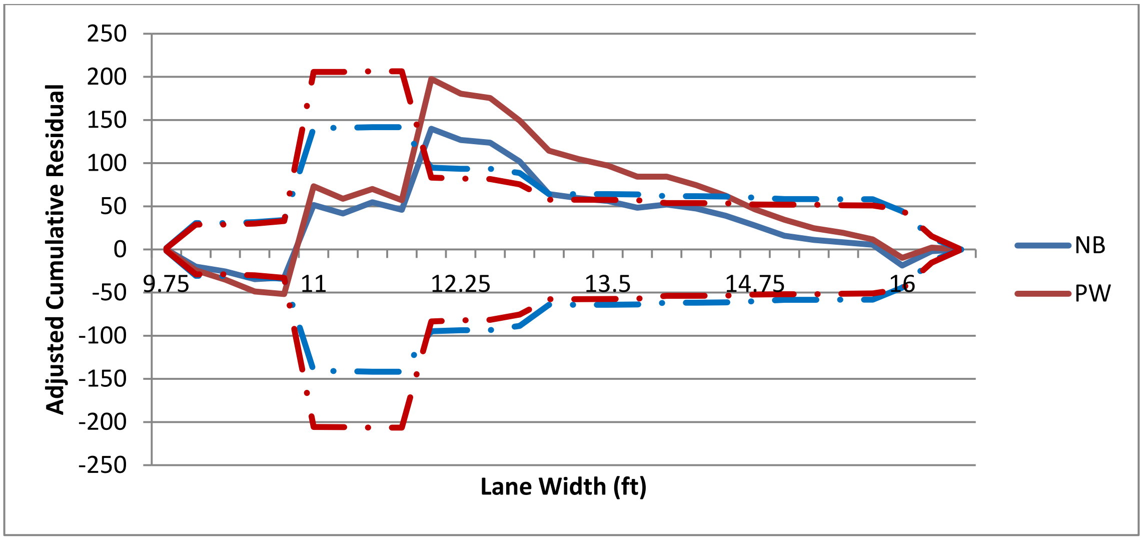
CURE PLOTS OF OTHER COVARIATES

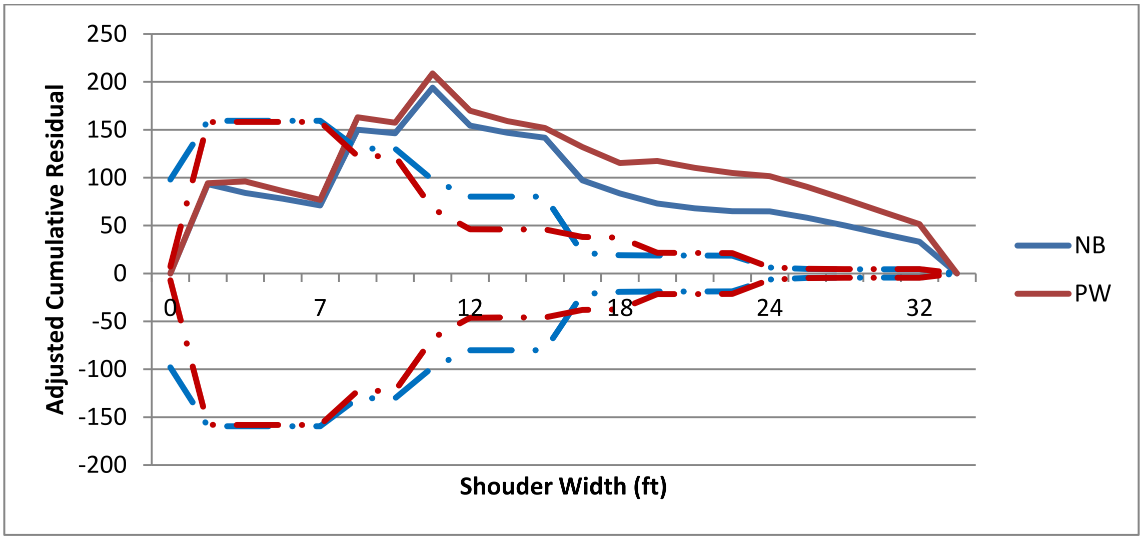
- **Indiana Data**

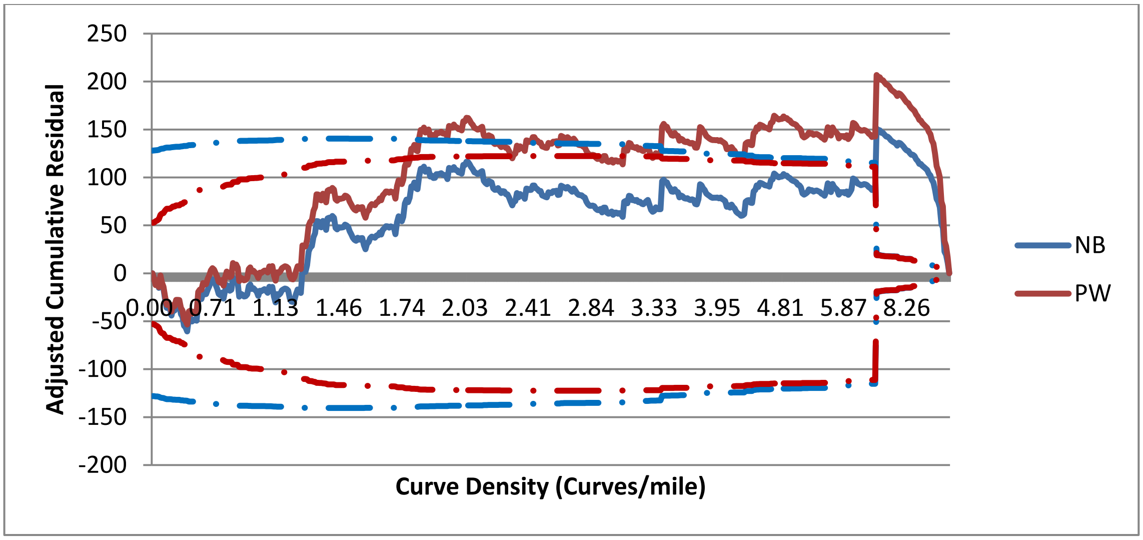




- **Texas Data**


250
200
150
100
50
0
-50
-100
-150
-200
-250
Adjusted Cumulative Residual
9.75
11
12.25
13.5
14.75
16
Lane Width (ft)
NB
PW


250
200
150
100
50
0
-50
-100
-150
-200
Adjusted Cumulative Residual
0
7
12
18
24
32
Shouder Width (ft)
NB
PW


250
200
150
100
50
0
-50
-100
-150
-200
Adjusted Cumulative Residual
0.00
0.71
1.13
1.46
1.74
2.03
2.41
2.84
3.33
3.95
4.81
5.87
8.26
Curve Density (Curves/mile)
NB
PW

VITA

Name: Lingzi Cheng

Address: Zachry Department of Civil Engineering, Texas A&M University,
College Station, Texas, 77843-3136.

Email Address: lingzi_cheng@hotmail.com

Education: M.S. Civil Engineering, Texas A&M University, 2012
B.S. Transportation Engineering, Wuhan University of Technology, 2010

Conferences: Cheng, L., S.R. Geedipally and D. Lord. “*Examining the Poisson-Weibull Generalized Linear Model for Analyzing Crash Data*”. Presented at the 91st Annual Meeting of the Transportation Research Board, Washington D.C., 2012.