EXAMINING THE USE OF REGRESSION MODELS FOR DEVELOPING CRASH MODIFICATION FACTORS

A Dissertation

by

LINGTAO WU

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Dominique Lord
Committee Members,	Jeffrey Hart
	Luca Quadrifoglio
	Yunlong Zhang
Head of Department,	Robin Autenrieth

May 2016

Major Subject: Civil Engineering

Copyright 2016 Lingtao Wu

ABSTRACT

Crash modification factors (CMFs) can be used to capture the safety effects of countermeasures and play significant roles in traffic safety management. The before-after study has been one of the most popular methods for developing CMFs. However, several drawbacks have limited its use for estimating high-quality CMFs. As an alternative, cross-sectional studies, specifically regression models, have been proposed and widely used for developing CMFs. However, the use of regression models for estimating CMFs has never been fully investigated. This study consequently sought to examine the conditions in which regression models could be used for such purpose.

CMFs for several variables and their dependence were assumed and used for generating random crash counts. CMFs were derived from regression models using the simulated data for various scenarios. The CMFs were then compared with the assumed true values. The findings of this study are summarized as follows: (1) The CMFs derived from regression models should be unbiased when the premise of cross-sectional studies were met (i.e., all segments were similar, proper functional forms, variables were independent, enough sample size, etc.). (2) Functional forms played important roles in developing reliable CMFs. When improper forms for some variables were used, the CMFs for these variables were biased, and the quality of CMFs for other variables could also be affected. Meanwhile, this might produce biased estimates for other parameters. In addition, variable correlation and distribution might potentially influence the CMFs and parameter estimates when improper functional forms were used. (3) Regression

models did suffer from the omitted-variable bias. If some factors having minor safety effects were omitted, the accuracy of estimated CMFs might still be acceptable. However, if some factors already known to have significant effects on crash risk were omitted, the estimated CMFs were generally unreliable. (4) When the influence on safety of considered variables were not independent, the CMFs produced from the commonly used regression models were biased. The bias was significantly correlated with the degree of their dependence.

DEDICATION

To my parents and wife for their love, support and encouragement.

ACKNOWLEDGEMENTS

First, I would like to express my deepest appreciation to my advisor, Dr. Dominique Lord, for his constant guidance, encouragement and support throughout this dissertation. My committee members, Dr. Jeffrey Hart, Dr. Luca Quadrifoglio, and Dr. Yunlong Zhang, are greatly appreciated for their suggestions on this dissertation. Dr. Bruce X. Wang is also appreciated for his comments.

I also want to extend my gratitude to my colleagues in Texas A&M Transportation Institute (TTI). Special thanks go to Mr. Robert Wunderlich, Dr. Troy Walden, Dr. David Bierling, Dr. Srinivas Geedipally, and Dr. Myunghoon Ko for their support on my research at TTI.

I would like to acknowledge International Road Federation and Research Institute of Highway, China for their financial support during my first year of Ph.D. study.

Finally, thanks to my parents, wife, and son for their patience.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

A crash modification factor (CMF) is a multiplicative factor that can be used to reflect or capture changes in the expected number of crashes when a given countermeasure or a modification in geometric and operational characteristics of a specific site is implemented (FHWA 2010; Gross et al. 2010; Wu et al. 2015). CMFs play a significant role in roadway safety management, including in safety effect evaluation, crash prediction, hotspot identification, countermeasure selection, and the evaluation of design exemptions. Several methods have been proposed for developing CMFs, such as the before-after (e.g., naïve or simple before-after, before-after with comparison group and empirical Bayes (EB) before-after), the cross-sectional (e.g., regression models and case-control), and expert panel studies among others (Gross et al. 2010). Amid these methods, before-after and cross-sectional studies are the most popular approaches (Shen and Gan 2003).

CMFs derived from before-after studies are based on the comparison of safety performance before and after the implementation of one or several treatments (or changes in the characteristics of the site(s)). Those derived from cross-sectional studies are based on the comparison in the safety performance of sites that have a specific feature with those that do not or are analyzed simultaneously based on datasets that contain a mixture of sites with different characteristics.

Over the last 15 years or so, the before-after study has been considered to be the best approach for developing CMFs (Gross and Donnell 2011; Gross et al. 2010). The

CMFs derived from before-after studies are usually believed to be more reliable than those produced from cross-sectional studies because it can directly account for changes that occurred at the sites investigated (Hauer 1997). However, although the before-after analysis is considered superior, high-quality CMFs derived from this approach is dependent on the availability of data (e.g., data availability for the before period, etc.) and the sample size (e.g., number of sites where the treatment of interest has been implemented, etc.). Furthermore, the estimated CMF can be biased if the regression-to-the-mean (RTM) and site selection effects are not properly accounted for in the before-after study (Davis 2000; Hauer 1997; Lord and Kuo 2012). Lord and Kuo (2012) even noted that the EB method can still be plagued by significant biases if the data collected for the treatment and control groups do not share the exact same characteristics.

Given the limitations of before-after studies described above, researchers have proposed that cross-sectional studies could be used for developing CMFs (Bonneson and Pratt 2010; Noland 2003; Tarko et al. 1999). Although different types of cross-sectional studies have been proposed over the years, the regression model (also known as safety performance function or SPF) remains the method of choice for estimating CMFs, as reflected by the large number of CMFs documented in the *Highway Safety Manual* (*HSM*) (AASHTO 2010) and Federal Highway Administration (FHWA) CMF Clearinghouse (FHWA 2010) that are derived from regression models.

Even though regression models are popular for developing CMFs, some researchers have criticized their use for such purpose because they may not properly

capture the relationships between crashes and variables influencing safety (Hauer 2005b, 2010, 2014). There are several assumptions with using regression models. For example, the primary premise of a cross-sectional study is that all locations are similar to each other for all factors affecting crash risk except those of interest (Gross et al. 2010). However, this requirement can hardly be satisfied in practice. In addition, some other problems (e.g., sample size, omitted variables, functional forms, independence assumption, etc.) also influence the modeling result, and hence the quality of CMFs produced from the regression analyses. Under such conditions, the CMFs derived from regression models may potentially be biased. In this context, so far, nobody has examined the statistical performance of CMFs that are developed using regression models. Thus, the primary objective of this study was to comprehensively investigate the robustness and accuracy of the CMFs derived from regression models. A secondary objective was to describe the conditions when the CMFs developed from regression models became unreliable and potentially biased. Note that, there has been some issues raised about whether or not cross-sectional studies are able to derive reliable cause-effect results, not only in traffic safety study, but also in other fields where this kind of statistical method has been used, such as psychology, epidemiology, etc. (Hauer 2015). The objective of this study was not to prove that the cross-sectional analyses are able to reveal the cause and effect of traffic collisions. Rather, the aim was to raise the potential problems associated with the commonly used regression models (i.e., generalized linear models or GLMs) for developing CMFs.

To accomplish the objectives, the following four tasks were conducted in this study:

Task 1 – Validation of CMFs derived from regression models

This task evaluated the accuracy of CMFs derived from regression models considering the most common and simple form (i.e., the linear relationship) and assuming all the assumptions of cross-sectional studies were satisfied. The purpose of this task was to validate whether or not the CMFs produced from regression models were reliable under ideal conditions.

Task 2 – Omitted variables and the accuracy of CMFs

In task 1, all factors that influenced crash risk were assumed to be known. But this requirement can hardly be met in reality. In this task, some factors affecting crash risk were assumed to be unknown or unable to be captured by the models. The purpose of this task was to investigate how the omitted-variable problem influenced the CMFs derived from regression models. More specifically, it was to quantify the problem.

Task 3 – Nonlinear relationships and the accuracy of CMFs

In tasks 1 and 2, the variables were assumed to have linear relationships (in the logarithmic form) with crash risk. This was consistent with the commonly used GLMs. However, some studies indicated that this may not be the case. Some variables had been shown to have nonlinear and/or non-monotonic relationships with crash risk (Gross et al. 2009; Hauer 2004). Under such conditions, the commonly used GLM method might not be applicable, and the CMFs produced from these models could be biased, especially

around the boundary areas. The purpose of this task was to evaluate the CMFs derived from regression models when some variables had nonlinear relationships with crash risk.

Task 4 – Combined safety effects and the accuracy of CMFs

Factors influencing safety are always assumed to be independent of each other when modeling crashes using the common methods. However, this may not be realistic in practice (Gross et al. 2010). It is common that multiple treatments were implemented at a problematic entity (e.g., a hotspot) simultaneously, and these treatments might have overlap effects on reducing crashes especially when the target collision types were the same. The CMFs derived from regression models might be biased if the variables were actually not independent. The purpose of this task was to investigate how this independence assumption influenced the accuracy of CMFs. This task is not about investigating the statistical correlation between variables, but the practical relationship and effects when multiple changes are applied simultaneously.

Each task contains one or multiple scenarios to assess the CMFs derived from regression models under various conditions.

This dissertation is divided into six chapters:

Chapter 2 documents the background about the commonly used approaches for developing CMFs, nonlinear relationships between variables and crash risk, and combined safety effects of multiple treatments.

Chapter 3 describes the methodologies used for estimating the quality of CMFs derived using regression models.

Chapter 4 provides detailed results of the simulation analyses.

Chapter 5 validates the findings based on observed data.

And finally, Chapter 6 summarizes the key findings of this study and provides avenues for further research.

2. BACKGROUND

This chapter provides relevant background pertaining to CMFs in three aspects: (1) the commonly used CMF estimating methods; (2) nonlinear relationships between variables and crash risk; and (3) the combined safety effects of multiple treatments.

2.1 Approaches for Estimating CMFs and Their Limitations

This section briefly describes the commonly used methods that have been proposed for estimating CMFs. The description mainly focuses on their advantages and limitations.

As mentioned above, before-after and cross-sectional studies are the two main approaches used to estimate CMFs. The CMF for a countermeasure derived from a before-after study is estimated by the change in the number of crashes occurring in a period before the improvement and the number occurring after the improvement (Gross et al. 2010; Shen and Gan 2003). Four types of before-after studies have been proposed to estimate CMFs: naïve before-after, before-after with comparison group, EB before-after and full Bayes (FB) before-after studies. The naïve before-after study simply assumes the crash performance before improvement is a good estimate of what would be in the after period if the countermeasure had not been implemented (Hauer 1997; Shen and Gan 2003). This approach is considered to be less reliable, because it does not account for changes unrelated to the countermeasure. The before-after study with

comparison group and EB before-after methods were then proposed to overcome this issue. Gross et al. (2010) documented the details of these CMF developing methods.

Even though multiple before-after studies have been developed and widely used to estimate CMFs, the comparison results from before-after studies may be inaccurate if the following issues are not properly accounted for:

Sample size – There might be inadequate samples of sites where the countermeasures of interest have been implemented. This will lead to statistical uncertainty (Gross and Donnell 2011).

RTM effect – This bias is related to the level of correlation for sites that are evaluated during different time periods. Sites that have large (or very small) values in one time period (say before) are expected to regress towards the mean in the subsequent period (Hauer 1997; Hauer and Persaud 1983).

Site selection bias – This is related to the RTM, but its effects are different in that the sites are selected based on a known or unknown entry criteria (e.g., five crashes per year). These entry criteria lead to a truncated distribution, which influences the before-after estimate (Lord and Kuo 2012).

Mixed safety effects – This bias or issue is related to when more than two or more countermeasures are simultaneously implemented at a roadway site, and there can be changes in traffic volume, weather, etc. after the implementation of treatments. This makes it difficult to evaluate the safety effect of a single countermeasure (Gross and Donnell 2011; Gross et al. 2010).

In contrast to before-after studies, cross-sectional studies compare the safety performance of a site or group of sites with the treatment of interest to similar sites without the treatment in a single point in time (Gross et al. 2010). The cross-sectional studies for developing CMFs can be regrouped into three categories: regression, case control and cohort methods. The regression method is currently the most frequently used approach because of its simplicity. It is usually accomplished through multiple variable regression models or SPFs. The SPFs can be used to quantify the effect of a specific variable on the predicted crash occurrence and CMFs are then derived from the model coefficients (Gross et al. 2010; Tarko et al. 1999).

Many models have been proposed to predict safety performance and hence to develop CMFs or crash modification functions (CM-Functions) (Lord and Mannering 2010). Although recent studies have introduced some new models for transportation safety analysis (Chen and Persaud 2014; Mannering and Bhat 2014; Park et al. 2014a; Zou et al. 2013a), the GLM with a negative binomial (NB) error structure is still the most popular method for modeling traffic crashes. Despite the fact that regression models have been extensively used in traffic safety studies, there are still some limitations with this approach:

Similarity in crash risk – A primary premise of a cross-sectional study is that all locations are similar to each other in all other factors affecting crash risk (Gross et al. 2010). However, this assumption seems to be unattainable in practice.

Omitted variables – A variety of variables can influence crash risk, but not all of them are measurable or can be captured in practice for model inclusion. It is common

that some SPFs were developed with limited variables, for example, using the traffic volume as the only variable in the model. This can lead to biased parameter estimates and incorrect CMFs (Lord and Mannering 2010).

Functional form – Functional form establishes the relationship between expected crashes and explanatory variables and is a critical part of the modeling process. Various forms have been used to link crashes to explanatory variables. But the modeling results tend to be inconsistent when using different functional forms (Miaou and Lord 2003). So far, there is no theory-based hypotheses to guide the choice of functional forms within regression models. Hauer (2015) pointed out it is a tall order to identify the right functional form.

There are also several other known issues with the crash modeling that will affect the CMFs derived from regression methods. Lord and Mannering (2010) and Mannering and Bhat (2014) provided more details of these issues.

Given the substantive issues associated with the before-after study and regression model method, it is not surprising that CMFs produced from these two approaches are not identical (Gross et al. 2013; Gross et al. 2010; Rodegerdts et al. 2007). For example, Hauer (1991) noted that the safety effects of some treatments tended to be different between those of cross-sectional and before-after studies. Further, the same approach and dataset can also generate different CMFs when using different regression models (Chen and Persaud 2014; Hauer 2010; Li et al. 2011; Lord and Bonneson 2007). Hauer (2010) illustrated the issue using a case of rail-highway grade crossing. A couple of previously conducted regression analyses and before-after studies about the safety effect

of rail-highway grade crossing were compared. The results of the former ones varied considerable and were obviously influenced by the choice of grouping method as well as choice of variables. On the other hand, the estimated effects of the six before-after studies were relatively consistent. Compared to the regression model, the before-after study has lower within-subject variability since it directly accounts for changes that have occurred at the study sites (Lord and Kuo 2012). Before-after studies are also less prone to confounding factors compared to cross-sectional studies (Carter et al. 2012). Furthermore, well-designed observational before-after studies provide advantages over other safety countermeasure evaluation methods (Gross and Donnell 2011). CMFs derived from regression models are suggested to be compared with those from before-after studies (Gross et al. 2010).

2.2 Nonlinear Relationships between Variables and Their Safety Effects

The coefficients are usually assumed to be fixed in the commonly used GLMs (i.e., the GLMs with linear additive link functions), and the CMF for a specific variable or treatment derived from the models is also fixed. This is, in fact, a linear relationship between the predicted crash risk and the changes in some variable (in the logarithmic form). The expected crash mean will always be multiplied by a constant factor when the variable increases by one unit, regardless of the original value of the variable. However, a fixed CMF may not properly account for the safety effects of the treatment on expected crash frequency because some variables may have nonlinear influences on crashes

(Hauer 2004; Hauer et al. 2004; Lee et al. 2015). Actually, some attempts have been made to explore the nonlinear effects.

Hauer et al. (2004) developed a statistical model to predict non-intersection crash frequency on urban four-lane undivided roadways. Several variables were considered in the analysis. Based on the estimated parameters, some variables were found to have linear or exponential influence on predicted crashes. However, some showed nonlinear effects on safety. For example, the degree of curve, which represented horizontal alignment, was captured to have a "U-shape" effect on on-the-road crashes. This indicated some flat curves might be safer than a tangent if this is true. But sharp curves would be associated with higher crash risk.

Xie and Zhang (2008) applied generalized additive models (GAMs) in traffic crash modeling. Compared to GLMs, GAMs used nonparametric smooth functions instead of parametric terms in GLMs, which made GAMs more flexible in modeling nonlinear relationships. Analysis result indicated GAMs performed better than GLMs in terms of goodness-of-fit (GOF) and predicting performance. This method was later utilized to develop CMFs for rural frontage segments in Texas (Li et al. 2011). The results showed that nonlinear relationships existed between crash risk and changes in lane and shoulder widths for frontage roads. For example, increasing shoulder width could bring relatively significant safety benefits when it was less than 6 ft. But when the shoulder width was between 6 and 8 ft, the CMF curve became flat, meaning widening shoulder had little influence on crashes. This result is slightly different with a previous GLM-based study (Lord and Bonneson 2007).

In order to capture the nonlinear relationships between variables and crashes, some neural network models have also been introduced into safety analysis. Xie et al. (2007) proposed Bayesian neural network (BNN) model for predicting motor vehicle crashes. BNN models had been previously reported to be able to effectively reduce the over-fitting phenomenon while still keeping the strong nonlinear approximation ability of neural networks (Xie et al. 2007). BNN models were estimated using the Texas frontage road data, the same used in several previous studies (Li et al. 2011; Lord and Bonneson 2007). Explicit functions between variables (e.g., lane width or shoulder width) and crash frequency were not available due to the black box property of BNN models. But the authors conducted sensitivity analysis of the trained BNN model for two sites. It was found that right shoulder width showed quadratic functions with predicted crash counts at the two sites, and lane width showed an "inverse U-shape" relation with crash counts at one site. Li et al. (2008) later conducted a continuation of this work. The researchers applied support vector machine (SVM) models to predict crashes, aiming to capture nonlinear relationships between explanatory and dependent variables. The Texas frontage road data was analyzed using SVM models and the results were quite similar with those using BNN method.

Recently, Lao et al. (2014) proposed generalized nonlinear models (GNMs) based approach to better elaborate non-monotonic relationships between variables and crash rates. Compared to GLMs, the major improvement of GNMs is using piecewise functions to capture the pattern between dependent and independent variables. This makes it more flexible to extract complex relationships between the two. Rear-end

crashes were modeled using GNM and GLM methods. Comparison showed GNMs outperformed GLMs. Meanwhile, some factors were found to be significant in GNMs but not in GLMs. Lee et al. (2015) later assessed the safety effects of changing lane width using GNMs. The main objective was to develop nonlinear relationships between lane width and crash rate. Various nonlinear link functions were used for the effects on crash rates of lane widths, and nonlinear CM-Functions were estimated for changing lane width. It was found that the CM-Function for lane width showed an "inverse U-shape" curve. It was combined with two quadratic functions and the 12-ft lane was found to be associated with the highest crash rates. This result contradicts some past studies, which concluded widening lanes could consistently reduce crash frequency (AASHTO 2010). More recently, Park and Abdel-Aty (2015a) assessed the safety effects of multiple roadside treatments (i.e., poles, trees, etc.) using GLM, GNM, and multivariate adaptive regression splines (MARS) model. The MARS model could capture both nonlinear relationships and interaction impacts between variables. Results showed that GNMs generally provided slightly better fits than the GLMs, and MARS model outperformed the other two. This indicated the roadside treatments had nonlinear effects on crash risk.

2.3 Safety Effects of Combined Treatments

A number of CMFs for various single treatments of roadway segments and intersections are provided in the *HSM*. No CMFs for combined treatments are available in the current version. However, it is common in practice that multiple countermeasures

are implemented simultaneously at a site to reduce the number and severity of collisions. The recommended approach (*HSM* method) of calculating the combined CMF for multiple treatments is multiplying the CMFs for individual elements or treatments together, as shown in Equation 2-1. Very limited combined safety effects have been reported in the CMF Clearinghouse (CMFClearinghouse 2014).

$$CMF_{comb} = CMF_{X_1} \times CMF_{X_2} \times \cdots \times CMF_{X_n} \quad (2\text{-}1)$$

Where,

CMF_{comb} = the combined CMF for *n* elements or treatments ($X_1, X_2, \cdots, X_n$); and,

CMF_{X_i} = the specific CMF for element or treatment X_i.

The main concept of this approach is that the simultaneously implemented treatments are independent. The safety effect of various countermeasures will not overlap when implemented at the same time. But this is not always true, especially when the target crashes of these countermeasures are the same. In such cases, the expected reduction in number of crashes will usually be lower than the sum of individual treatments. And the product of individual CMFs will underestimate the true combined CMF (i.e., safety benefits are overestimated) (Bonneson and Lord 2005; Harkey et al. 2008; Roberts and Turner 2007). To address this problem, researchers have proposed a couple of alternatives for estimating combined effects of multiple treatments, e.g., reducing the safety effects of less effective treatments, applying only the most effective CMF, multiplying weighted factor (Turner method), weighted average of multiple CMFs (also known as meta-analysis method), etc. More details of these methods are documented in Gross et al. (2012), Elvik (2009) and Gross and Hamidi (2011). A

common concept within these approaches is that simultaneously implemented treatments usually have overlapped safety effects.

Park et al. (2014b) estimated CMFs for two single treatments (installing shoulder rumble strips, and widening shoulder width) and the combined CMF for implementing the two simultaneously on rural multi-lane highways. The results confirmed that the combined CMFs, in general, did not equal to the product of the two single CMFs. The researchers further calculated CMFs for multiple treatments using various combining methods and compared them with those estimated using real data. It was found that each method applied to different crash types and injury levels.

Park and Abdel-Aty (2015b) later developed adjustment functions for combined CMFs. An adjustment factor (AF) or adjustment function (A-Function) was introduced to assess the combined safety effects of two treatments (installing shoulder rumble strips, and widening shoulder width) on rural two-lane highways. An AF higher than 1.0 indicated the combined amount of crash reduction was lower than the sum of individual treatments. And vice versa if it was less than 1.0. Particularly, when it equaled to 1.0, the treatments were independent of each other. The AF (or A-Function) used in the study is shown in Equation 2-2.

$$CMF_{comb} = CMF_{X_1} \times CMF_{X_2} \times \cdots \times CMF_{X_n} \times AF \quad (2\text{-}2)$$

Where,

AF = the adjustment factor for treatments $X_1, X_2, \cdots, X_n$, $AF > 0$.

Three nonlinear A-Functions for the combined CMFs were developed considering different crash types and severities. All of them were higher than 1.0, which

indicated the combined CMFs calculated using *HSM* method were underestimated. The amount of underestimation varied based on crash types and severities. In addition, the AFs also varied as the original shoulder width changed rather than kept as constant values. That means the level of dependence between the two treatments was not identical among all conditions.

Although only a few studies estimated the combined effects of multiple safety treatments (Bauer and Harwood 2013; De Pauw et al. 2014; Park and Abdel-Aty 2015b; Park et al. 2014b; Wang et al. 2015), it has shown that some treatments or highway characteristics do influence crashes dependently. Under such conditions, the independence assumption of regression models cannot be met. This might potentially reduce the quality of CMFs. No matter which CMF combination method is used, reliable individual CMFs are critical for estimating safety effects of both combined and single treatments.

2.4 Summary

The primary findings from the literature review are summarized below:

(1) Both before-after and cross-sectional studies have their own drawbacks. So far, no study has fully investigated whether or not the CMFs derived from regression models really reflect the true safety effects of treatments. It is necessary to evaluate the accuracy of CMFs estimated from regression models.

(2) In the previous studies, analyses using nonlinear methods generally showed better results than the commonly used GLM approach. This indicates some variables

indeed have nonlinear and/or non-monotonic effects on crash frequency, and the CMFs derived using normal GLMs may not be able to adequately capture these types of relationships.

(3) The combined CMF of multiple treatments do not always equal to the productive of single CMFs of individual treatments. In other words, some treatments or highway characteristics are not actually independent. That is to say the independence assumption of regression models cannot always be met in practice. This might potentially reduce the quality of CMFs. It is necessary to examine the accuracy of individual CMFs derived from regression models considering the dependence of variables.

This chapter has introduced some relevant background of CMFs/CM-Functions and numbers of potential problems with the common CMF developing approaches. The next chapter documents the methodology used to evaluate the quality of CMFs.

3. METHODOLOGY

This chapter describes the methodologies regarding how to examine the accuracy of CMFs derived from regression models. Sections 3.1 provides the simulation protocol for estimating the accuracy of CMFs. Section 3.2 describes the methodology for quantifying the omitted-variable problem. Section 3.3 mainly introduces the measurement used to quantify the nonlinearity. Sections 3.4 considers variables correlations, a common phenomenon with practical crash data. Section 3.5 presents the specific methodology to investigate the independence problem. Section 3.6 describes the simulated datasets. And finally, Section 3.7 summarizes all the scenarios in this study.

3.1 Simulation Analysis for Linear Relationships

This section first describes the simulation protocol used to estimate the accuracy of CMFs derived from regression models, mainly focusing on linear relationships. Following that, a simulation example is provided to illustrate the specific procedures. And the scenarios with linear relationship are summarized in the last part.

3.1.1 Simulation Protocol

To investigate the use of regression models for developing CMFs, CMFs for different variables have to be derived from regression models and compared with their true safety effect. However, the exact safety effect of a feature or treatment is hardly known in the real world, this makes it extremely difficult to examine the CMFs when observed crash data are used. But, by analyzing simulated data, one can compare the

CMFs estimated from regression models with the assumed true values. So, this study mainly used simulated data.

This section establishes a simulation protocol for evaluating CMFs derived from regression models. The simulation experiment used in this study was proposed by Hauer (2014). First, CMFs (i.e., safety effects) for some highway geometric features were assumed. Then, random crash counts were simulated based on the assigned values of CMFs. Finally, the CMFs were estimated from the simulated crash data and compared with the true CMFs. This research adopted this simulation procedure, but necessary changes were made. The simulation contains five steps, as described in detail below:

Step 1: *Assign initial values*

Assume CMFs for highway geometric features of interest. Tasks 1 and 2 (i.e., linear relationship and omitted variable) assumed an exponential relationship between a highway geometric feature and its safety effect. For example, it was assumed that the CMF for lane width was $CMF_{LW\_Assumed}$, meaning the expected crash frequency was multiplied or divided by $CMF_{LW\_Assumed}$ if the lane width increased or decreased by one foot. Task 3 assumed multiple forms of relationships (i.e., linear and nonlinear) between variables and crash risk.

Step 2: *Calculate mean values*

Calculate the true crash means for each segment using SPFs and assumed CMFs using Equation 3-1 (AASHTO 2010).

$$N_{true,i} = N_{spf,i} \times (CMF_{1,i} \times CMF_{2,i} \times \cdots \times CMF_{m,i}) \times C \quad (3\text{-}1)$$

Where,

$N_{true,i}$ = true crash mean for roadway segment *i* for a certain time period (i.e., one year). The true crash mean was the theoretical number of crashes that occur on a segment, it was used to generate random crash counts in this study;

$N_{spf,i}$ = crash mean for roadway segment i for the base conditions, generated from an SPF;

$CMF_{j,i}$ = assumed CMF specific to geometric feature type j of segment i, $j = 1, 2, \ldots, m$;

m = the total number of variables or geometric features of interest; and

C = calibration factor to adjust SPF for local conditions, and was assumed to be 1.0 for all segments in this study.

The SPF in this study was adopted from the *HSM (AASHTO 2010)* for rural two-lane highways (the same as the data used in this study, described in Section 3.6), as shown in Equation 3-2.

$$N_{spf,i} = AADT_i \times L_i \times 365 \times 10^{-6} \times e^{-0.312} = 2.67 \times 10^{-4} \times L_i \times AADT_i \quad (3\text{-}2)$$

Where,

$AADT_i$ = average annual daily traffic (AADT) volume (vehicles per day) of segment *i*; and

L_i = length of roadway segment *i*, (mile).

Step 3: *Generate discrete counts*

Generate random counts Y_i given that the mean for segment i was gamma distributed with dispersion parameter α (the inverse dispersion parameter, $\phi = 1/\alpha$) and mean equal to 1 (Lord 2006):

$$\mu_i = N_{true,i} \times exp(\varepsilon_i) \quad (3\text{-}3a)$$

$$exp(\varepsilon_i) \sim Gamma(1,\ \alpha) \quad (3\text{-}3b)$$

$$Y_i \sim Poisson(\mu_i) \quad (3\text{-}3c)$$

Where,

μ_i = Poisson mean for segment *i* for a certain time period;

ε_i = model error independent of all the covariates, and $exp(\varepsilon_i)$ was assumed to be independent and gamma distributed with mean equal to 1 and dispersion parameter equal to α ; and,

Y_i = randomly generated crash counts for segment i for a certain time period.

Thus, the simulated crash counts followed Poisson-Gamma or NB distribution with parameters ϕ and μ_i. The probability density function (PDF) is given by Equation 3-4 (Lord 2006).

$$f(y_i;\phi,\mu_i) = \frac{\Gamma(y_i+\phi)}{\Gamma(\phi)y_i!}\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi}\left(\frac{\mu_i}{\phi+\mu_i}\right)^{y_i} \quad (3\text{-}4)$$

Where,

y_i = crash count for segment *i* for a certain time period;

μ_i = the crash mean during a period for segment *i*; and,

ϕ = inverse dispersion parameter.

Step 4: *Estimate CMFs from the simulated crash data*

As has been documented in the background, many models and functional forms have been proposed to predict crashes. In this study, the most commonly used GLM and functional form were selected, as shown in Equation 3-5 (Lord and Bonneson 2007). Note that a different parameter for describing the mean of the site, Λ_i, was used for estimating the models (compared to the one used for the simulation, μ_i).

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\sum_{j=2}^{n} \beta_j \times x_j) \quad (3\text{-}5)$$

Where,

$E(\Lambda_i)$ = the estimated crash mean during a period for segment *i*;

x_j = a series of variables, such as the lane width of segment *i*; and,

$\beta_0, \beta_1, \ldots, \beta_n$ = coefficients to be estimated.

For the GOF of the models, the following three methods were used: (1) Akaike information criterion (AIC), (2) Mean absolute deviance (MAD), and (3) Mean-squared predictive error (MSPE). More information about MAD and MSPE are documented in Lord et al. (2008).

Once the model was fitted and coefficients were estimated using the simulated crash data, the CM-Function for variable *j* was then derived as (Gross et al. 2010; Lord and Bonneson 2007):

$$CMF_{x,j} = exp[\beta_j \times (x - x_{0,j})] \quad (3\text{-}6)$$

Where,

β_j = estimated coefficient for variable *j*;

x = value of variable *j*, such as lane width, curve density;

$x_{0,j}$ = base condition defined for variable *j*, usually 12 ft for lane width; and,

$CMF_{x,j}$ = CMF specific to variable *j* with value of *x*.

This also indicated the CMF derived from the SPF for variable *j* was $CMF_j = exp(\beta_j)$, meaning the expected crash frequency would be multiplied or divided by CMF_j if the variable *j* increased or decreased by one unit.

Repeat Steps 2 to 4 100 times, calculate the mean and the standard deviation of the estimated CMF values for each variable.

Step 5: *Evaluate the CMF derived from the regression models*

Two indexes, estimation bias and error percentage, were used to evaluate the CMF derived from SPFs. They are shown in Equations 3-7 and 3-8. The smaller is the error percentage, the more accurate the CMF derived from SPFs is.

$$\Delta_j = CMF_{j\_Assumed} - CMF_{j\_SPF} \quad (3\text{-}7)$$

$$e_j = 100 \times \frac{|\Delta_j|}{CMF_{j\_Assumed}} \quad (3\text{-}8)$$

Where,

Δ_j = estimation bias of CMF for variable *j*;

e_j = error percentage of CMF for variable *j*, (%);

$CMF_{j\_Assumed}$ = assumed CMF value for variable *j*; and

CMF_{j_SPF} = CMF derived from the SPF for variable *j*.

Please note the meaning of terminology "bias" used above to quantify the quality of CMFs. In Mathematics and Statistics, bias is defined as a systematic (built-in) error which makes all values or estimates wrong in the same direction and by a certain amount (Math is Fun 2014). Specifically, bias in this dissertation means the difference between the true CMF for a variable and that estimated from regression models. It can also be defined as misspecification error (as some CMFs are misestimated in the models). However, to simplify the description, the issue of misspecification is referred as "bias" in the rest of this dissertation.

3.1.2 Simulation Example

This section provides an example to illustrate the various steps used for generating crash data and the method of evaluating CMFs.

Table 1 below shows snapshots of the simulated crash counts for *N* segments considering a single variable of lane width. In this table, the CMF was assumed to be 0.9, which meant the increase of one foot in lane width decreased the predicted number of crashes by 10 percent (1.0 - 0.9), with the base condition for a lane width equal to 12 ft. So, the CMF specific to a segment *i* can be calculated as Equation 3-9. The assumed CM-Function for lane width is shown in Figure 1 (the dotted line with triangles).

Table 1 Example of Simulated Crash Counts for N Segments

Seg.	L [a]	AADT	LW [b]	CMF	N_{spf}	N_{true}	Yr 1	Yr 2	Yr 3
1	0.113	15360	11	1.111	0.47	0.52	0	0	1
2	0.213	18420	8	1.524	1.05	1.60	1	3	2
3	0.125	4260	9	1.372	0.14	0.20	0	0	1
4	0.161	10600	10	1.235	0.46	0.56	1	0	1
5	0.196	12560	12	1.000	0.66	0.66	1	0	2
...	...	...	...	...	...	...	...	...	...
N	0.234	4580	13	0.900	0.29	0.26	0	1	1

Note: a – L = length (mile); b - LW = lane width (ft).

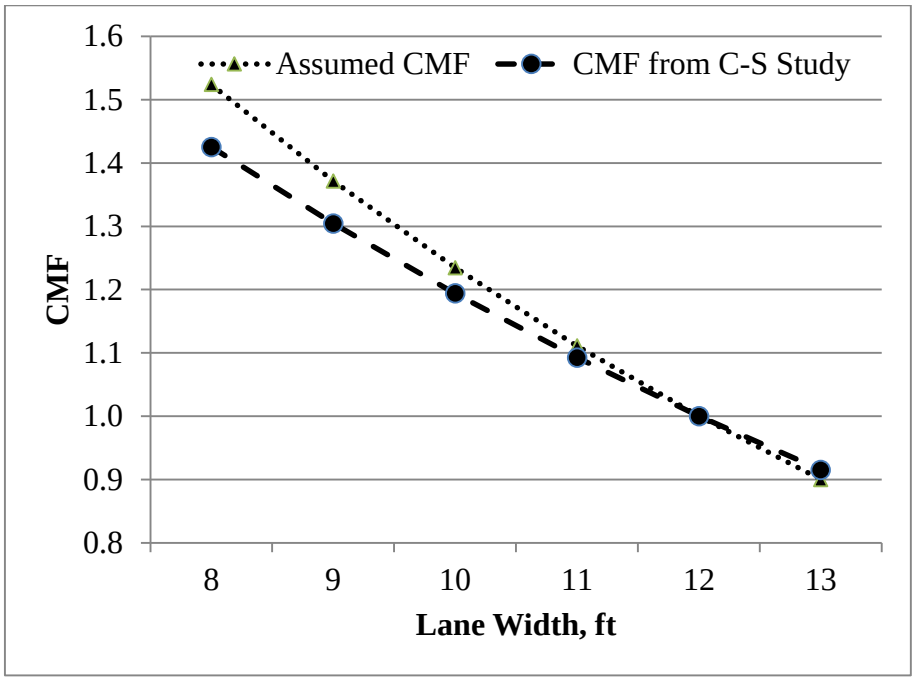


Figure 1 An Example Illustrating the Assumed CMF and CMF Derived from SPFs

The CMF for lane width is given by:

$$CMF_{LW_i,i} = 0.9^{LW_i - 12} \quad (3\text{-}9)$$

Where,

LW_i = lane width of segment *i* (ft); and

$CMF_{LW_i,i}$ = specific CMF for lane width of segment *i*.

Thus, the true crash mean of segment *i* was calculated as (recall that the calibration factor was assumed to be 1.0 for all segments in this study):

$$N_{true,i} = N_{spf,i} \times CMF_{LW,i} \quad (3\text{-}10)$$

Then, the $exp(\varepsilon_i)$ of each segment was randomly generated based on a Gamma distribution with parameters mean equal to 1 and dispersion parameter equal to α, which had the value of 2 in Table 1. μ_i of segment *i* was then calculated by multiplying $N_{true,i}$ and $exp(\varepsilon_i)$, as shown in Equation 3-3b.

After, a sequence of Poisson counts were generated based on the mean μ for each segment. Three years of simulated crash counts are shown in the last three columns in Table 1. The theoretical function form of these crash counts is shown in Equation 3-11.

$$N_{true,i} = N_{spf,i} \times CMF_{LW,i} = 2.67 \times 10^{-4} \times L_i \times AADT_i \times 0.9^{LW_i - 12} \quad (3\text{-}11a)$$

Or equivalently,

$$N_{true,i} = 9.45 \times 10^{-4} \times L_i \times AADT_i \times exp(-0.105 \times LW_i) \quad (3\text{-}11b)$$

The simulated crash data was analyzed using the NB regression model. The mean functional form is provided in Equation 3-12.

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\beta_2 \times LW_i) \quad (3\text{-}12)$$

The coefficients of Equation 3-12 were estimated using a NB regression model in MASS package (Ripley et al. 2014) within the software R. The GOF measures were calculated using Metrics package (Hamner 2013). The modeling output is shown in Table 2. The p-values indicate the variables were statistically significant at the 99 percent level in this example. And, the small MAD and MSPE show the modeling result performed well (given the simulated data).

Based on the fitting result, the CM-Function for lane width derived from this SPF is shown in Equation 3-13.

$$CMF_{LW} = exp[\beta_2 \times (LW - 12)] = 0.915^{LW-12} \quad (3\text{-}13)$$

The value of 12 in Equation 3-13 reflects the base condition for lane width, which means that the CMF derived from prediction model is equal to 1.0. In this case, with an increment of one foot in lane width, the crash mean was expected to be multiplied by $e^{\beta_2} \approx 0.915$. So, the CMF derived from the SPF was 0.915 in this example. The CM-Function is also shown in Figure 1 (the dash line with circles).

Table 2 Modeling Output of the Example Data

Model Variable	Theo. Value [a]	Coef. Value [b]	SE [c]	p-Value
Intercept [$ln(\beta_0)$]	$ln(9.45\times10^{-4})$ =-6.96	-6.810	0.340	5.3911E-89
Ln(AADT) (β_1)	1.00	0.960	0.036	9.996E-161
Lane Width (β_2), ft	-0.105	-0.089	0.012	1.5953E-13
AIC			20614.5	
MAD			0.214	
MSPE			0.244	

Note: a – theoretical value; b – estimated coefficient value; c – SE = standard error.

The bias between the assumed CMF and that from the SPF (without repeat in this example) is calculated as:

$$\Delta = CMF_{Assumed} - CMF_{SPF} = 0.90 - 0.915 = -0.015$$

And the error percentage is:

$$e = 100 \times \frac{|bias|}{CMF_{Assumed}} = 100 \times \frac{|-0.015|}{0.90} = 1.69(\%)$$

By repeating the Steps 2 to 4 100 times, 100 CMFs could be estimated. The mean and standard deviation of CMFs and mean of GOF measures could be calculated. The estimation bias and error percentage were then calculated based upon the mean value of derived CMFs. For illustration purposes, this example only considered one variable, lane width.

3.1.3 Scenarios

Two scenarios were examined in Task 1 to accommodate more complex situations with different levels of dispersions (i.e., inverse dispersion parameters) and two additional variables, curve density and pavement friction. The scenarios are described below. The scenarios were names as "Scenario Number". To make it consistent, the scenarios in the following tasks were given similar names, and the scenario numbers were continuous.

Scenario I: Consider one variable only, linear relationship

Various CMF values were assumed for lane width in this scenario. The objective was to examine whether or not the regression models can produce reliable CMFs when all the requirements of a cross-sectional study were satisfied.

Scenario II: Consider three variables, linear relationship

This scenario considered three variables, lane width, curve density and pavement friction. Details about the last two variables are introduced in Section 3.6.1. A fixed CMF value was assigned for each of the three variables. The objective was to examine whether the CMFs derived from SPFs were reliable when multiple variables were considered.

To reflect different traffic characteristics, the inverse dispersion parameter ϕ in the two scenarios varied between 0.5, 1.0 and 2.0, respectively.

3.2 Omitted Variable Problem

As has been documented in Chapter 2, omitted variable is an important problem with regression models. This problem can lead to biased parameter estimates in the regression models and incorrect CMFs. This task investigated how the omitted-variable problem influenced the CMFs derived from regression models. In the previous section, all factors that influenced crash risk were assumed to be known. In contrast, not all the factors affecting crash risk were known or able to be captured by the model in this section. One scenario (i.e., Scenario III) was studied to address this problem, as described below.

Scenario III: Omitted variables, linear relationship

This scenario considered three variables, lane width, curve density and pavement friction. Their CMFs were assumed to be in linear forms, the same as that in Scenario II. But only one variable, the lane width, was included in the SPF; this fell under the

omitted-variable problem. The inverse dispersion parameter ϕ in this scenarios also varied between 0.5, 1.0 and 2.0, respectively.

The methodology used to evaluate the quality of CMFs in this section was essentially the same as that in Section 3.1, except that the two variables were excluded in the regression model.

3.3 Nonlinear Relationships

This section describes how the accuracy of CMFs derived from SPFs was investigated when some variables had nonlinear relationships. Intuitively, if the nonlinear relationship is weak (the CM-Function curve is quite flat or approximately a straight line), the accuracy of CMFs derived from SPFs should be similar to those in the previous scenarios. In the contrast, if the nonlinear relationship is strong (the curve is sharp), the accuracy of CMFs may be potentially affected. A measurement is necessary to describe how flat or sharp the curve is. This section first introduces the concept of quantifying nonlinearity, then presents the scenarios.

3.3.1 Quantifying Nonlinearity

First, the definition of ***the closest line to a curve***. For a given integrable curve $y = f(x)$ over $[m, n]$, the closest line to this curve is defined as a straight line $y = k \times x + c$ that minimizes the area between the two. This definition is illustrated in Figure 2. The dashed curve represents the given function $y = f(x)$, and the solid line represents the closest line to this curve $y = k \times x + c$. This line minimizes the area

between the two (the shadowed area in Figure 2). Given the range, in general, the larger the area is, the stronger the nonlinearity the curve tends to have. Particularly, if the given function is linear, the closest line is the function itself, and the area is technically equal to zero.

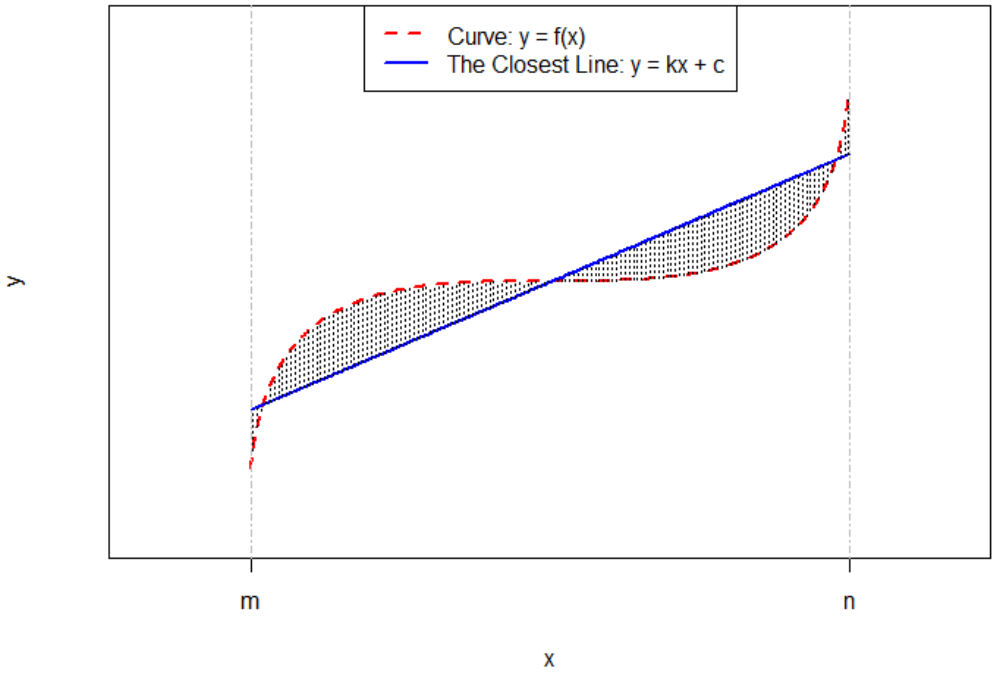


Figure 2 Example Illustrating the Closest Line to a Curve

Second, the definition of ***average vertical distance between a curve and its closest line***. Although the area between a curve and its closest line can reflect the nonlinearity of the curve, the area still depends on the range. Wider range is more likely to yield larger area. The variables affecting traffic crashes are usually different in their possible values in practice. For example, the lane width may vary from 8 ft to 13 ft,

while the curve density may vary from 0 to 16 curves per mile. A standardized measurement is necessary to quantify the nonlinearity. The average vertical distance (AVD) between a curve and its closest line is defined as the area between the two divided by the range. So, in Figure 2, the AVD is calculated as dividing the shadowed area by $n-m$. This way, the AVD itself can be used to quantify the nonlinearity of a curve regardless of its range. The larger this distance is, the stronger the nonlinearity that curve has. If the given function is linear, the AVD is zero.

The details for calculating the coefficients of the line (i.e., k and c) and AVD are shown below. The objective is to minimize the area, shown in Equation 3-14.

$$Area = \int_m^n | f(x) - (k \times x + c) | dx \tag{3-14a}$$

Or equivalently,

$$Z = \int_m^n [f(x) - (k \times x + c)]^2 dx \tag{3-14b}$$

k and *c* can be easily derived through mathematical translations, shown below.

$$k = \frac{(n-m)\int_m^n xf(x)dx - \int_m^n f(x)dx \int_m^n xdx}{(n-m)\int_m^n x^2 dx - [\int_m^n xdx]^2}$$

$$c = \frac{\int_m^n f(x)dx \int_m^n x^2 dx - \int_m^n f(x)dx \int_m^n xdx}{(n-m)\int_m^n x^2 dx - [\int_m^n xdx]^2}$$

The area can be calculated by substituting *k* and *c* into Equation 3-14a, and the AVD is then calculated as dividing the area by $n-m$. The AVD was used to measure the nonlinearity of a CM-Function in this study.

3.3.2 Scenarios

The method used to evaluate the accuracy of CMFs with nonlinear relationships with crash risk was similar with that of linear relationships. The only difference was that nonlinear CM-Functions were assumed for some variable(s) and used to generate crash counts. Three scenarios were analyzed in this section, as described below.

Scenario IV: Consider one variable only, in nonlinear form

Only lane width was considered and assumed to have nonlinear effects on crash. The main objective was to examine the bias of CMF for a variable with different levels of nonlinearity.

Scenario V: Consider three variables, only one in nonlinear form

Three variables, lane width, curve density and pavement friction, were considered in this scenario. Fixed CMFs were assigned to curve density and pavement friction. The CM-Functions for lane width were assumed to be in nonlinear forms. The main objective was to examine the influence of nonlinear variables on the accuracy of CMFs for linear variables.

Scenario VI: Consider three variables, two in nonlinear form

This scenario was similar with Scenario V, but both lane width and curve density were assumed to have nonlinear relationships. The CMF for pavement friction was fixed. The main objective was to examine the influence of nonlinear variables on the accuracy of CMFs for both linear and nonlinear variables.

For all the three scenarios, the assumed nonlinear relationships varied from weak to strong. Thus, each scenario contained a number of sub-scenarios. In addition, the inverse dispersion parameter ϕ in each sub-scenario varied between 0.5, 1.0 and 2.0.

Note that, in Section 3.1, the theoretical functions for generating crash counts and the considered functional forms in regression models were of the same family (i.e., linear form or relationship). And the assumed CMFs and those generated from SPFs were similar (i.e., single ones). Only one bias and error percentage were calculated to examine the quality of CMFs. However, in this section, the theoretical functions were in nonlinear forms, which were not of the same family as the considered functional forms. No single CMF could be used to represent the assumed CM-Function. Thus, one bias or error percentage was not enough to assess the quality of CMFs. To overcome this problem, several specific CMFs for variables at typical points from both the assumed CM-Function and that developed from regression models were compared, and the bias and error percentage were calculated based on the specific CMFs.

3.4 Variable Correlation

In the previous sections, all the considered variables were assumed to be (perfectly) independent of each other, and each was uniformly or discrete uniformly distributed among the corresponding range (as will be shown in Section 3.6.1). However, this might not be the case in practice. Some variables may be highly correlated with each other. For example, when constructing two highways, one with higher demand (i.e., AADT) and the other with lower, it is common that the former one will be designed

with higher standard, e.g., wider lanes and shoulders, etc. Thus, variables AADT and lane width are correlated. And also, in highway design manuals (AASHTO 2004; TxDOT 2014), lane width is recommended to be 12 ft for most highways. So 12 ft may be prevalent among lanes, and it is not discrete uniformly distributed in practices. This might affect the regression result and hence the CMFs for variables. This section aimed to examine whether or not variable correlation had influence on the CMFs derived from regression models. Two scenarios (i.e., Scenario VII and VIII) were analyzed, as described below.

Scenario VII: Variable correlation, linear relationship

This scenario considered one variables, lane width, only. Various CMF values were assumed for lane width. This scenario was basically the same as Scenario I, except the variable lane width was correlated with AADT.

Scenario VIII: Variable correlation, nonlinear relationship

Only lane width was considered and assumed to have nonlinear effects on crash. This scenario was basically the same as Scenario V, except the variable lane width was correlated with AADT.

The inverse dispersion parameter ϕ in the two scenarios varied between 0.5, 1.0 and 2.0, respectively. The methodology used to evaluate the quality of CMFs in this section was also essentially the same as that in Sections 3.1 and 3.3, except that two variables AADT and lane width were correlated. A new dataset was generated, as will be described in Section 3.6.2.

3.5 Combined Safety Effect

A similar approach was used to evaluate the CMFs derived from regression models considering the dependence of variables (not in a statistical sense), but it was modified to fit the specific characteristics of this task. The main concepts were: (1) assume CMFs and dependence for variables; (2) generate random crash counts; and (3) estimate CMFs using regression models and compare them with the assumed true values.

The major difference in this section was the use of adjustment factors (AFs). An adjustment factor was assumed to capture the combination effect of multiple treatments. This was similar to the method used in the recent study by Park and Abdel-Aty (2015b). The combined CMF for multiple treatments is calculated by Equation 3-15.

$$CMF_{comb} = CMF_{X_1} \times \cdots \times CMF_{X_n} \times AF^{I_{\{X_1 \neq X_{1base}\}}(X_1) \times \cdots \times I_{\{X_n \neq X_{nbase}\}}(X_n)} \quad (3\text{-}15)$$

Where,

CMF_{comb} = the combined CMF for a segment;

CMF_{X_j} = the assumed specific CMF for variable X_j of the segment;

AF = assumed adjustment factor for variables X_1, X_2, $\cdots$, X_n;

X_{jbase} = the base condition for variable X_j; and,

$I_{\{X_j \neq X_{jbase}\}}(X_j)$ = indicator function for variable X_j. It equaled to zero if variable X_j of the segment was equal to the base condition, otherwise 1.0.

The indicator functions made the adjustment factor to be working or not based on specific conditions of the segment and the presumed dependence relationships between variables.

To simplify the analysis, only two variables, lane width and shoulder width, were considered in this scenario (i.e., Scenario IX). And each variable in the dataset was assigned one of two values: the baseline and improved, respectively. For lane width, it was either 12 ft (baseline) or 13 ft (wider lane). And for shoulder width, it was either 6 ft (baseline) or 7 ft (wider shoulder). This way, the total segments could be classified into four categories: (1) baseline; (2) wider lane; (3) wider shoulder; and (4) wider lane and wider shoulder. They are described in Table 3.

The CMF for lane width was assumed to be CMF_{LW} with baseline equal to 12 ft. So, the specific CMFs for lane widths of 12 ft and 13 ft were 1.0 and CMF_{LW}, respectively. Similarly, the CMF for shoulder width was assumed to be CMF_{SW} with baseline equal to 6 ft. The specific CMFs for shoulder widths of 6 ft and 7 ft were 1.0 and CMF_{SW}, respectively. This study assumed neither CMF_{LW} nor CMF_{SW} equaled to 1.0. Furthermore, the adjustment factor was used to capture the dependence of the safety effects of the two variables. That is to say, if a segment was wider in both lane and shoulder, the combined CMF was multiplied by the adjustment factor. The CMFs for lane width, shoulder width and combined CMF for the four groups of segments are shown in the last three columns of Table 3.

Table 3 Summary of Four Groups of Segments

Group	LW (ft)	SW (ft)	CMF for LW	CMF for SW	Combined CMF
Baseline	12	6	1.0	1.0	1.0
Wider Lane	13	6	CMF_{LW}	1.0	CMF_{LW}
Wider Shoulder	12	7	1.0	CMF_{SW}	CMF_{SW}
Wider Lane and Wider Shoulder	13	7	CMF_{LW}	CMF_{SW}	$CMF_{LW} \times CMF_{SW} \times AF$

Note: LW = lane width; SW = shoulder width.

Specifically, the assumed CMF for lane width (i.e., CMF_{LW}) varied between 0.8 and 0.9. And that for shoulder width (i.e., CMF_{SW}) varied between 0.85 and 0.9. The adjustment factor changed from 0.80, 0.90, 0.95, 1.05, 1.10 to 1.20. When the adjustment factor is less than 1.0, it means widening both lane and shoulder width simultaneously will bring more safety benefits than the "sum" of the two single treatments. The smaller the adjustment factor is, the more benefit will be. In contrast, if it is more than 1.0, taking the two treatments simultaneously will have a lower effect than their "sum." The higher the adjustment factor is, the lower the combined safety effect will be.

In total, there were 24 sub-scenarios in this section, shown in Table 4. The inverse dispersion parameter (ϕ) varied between 0.5, 1.0 and 2.0 in each sub-scenario to reflect different traffic characteristics.

The theoretical function of the generated crash counts in this scenario is shown in Equation 3-16.

$$N_{true,i} = N_{spf,i} \times CMF_{comb,i} = 2.67 \times 10^{-4} \times L_i \times AADT_i \times CMF_{comb,i} \quad (3\text{-}16)$$

Where,

$N_{true,i}$ = true crash mean for roadway segment *i* during a certain time period (i.e., one year);

$AADT_i$ = AADT of segment *i* (vehicles per day);

L_i = length of segment *i* (mile); and,

$CMF_{comb,i}$ = the combined CMF for lane width and shoulder width of segment *i*. It was calculated by the methods shown in Table 3 (the last column).

The CMFs for the two variables were derived from SPFs with similar procedures utilized in the previous study. The considered functional form is shown in Equation 3-17, in which the two variables (i.e., lane width and shoulder width) were assumed to influence crashes independently.

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\beta_2 \times LW_i + \beta_3 \times SW_i) \quad (3\text{-}17)$$

Where,

$E(\Lambda_i)$ = the estimated crash mean during a period (i.e., one year) for segment *i*;

LW_i = lane width of segment *i* (ft);

SW_i = shoulder width of segment *i* (ft); and

$\beta_0, \beta_1, \beta_2, \beta_3$ = coefficients to be estimated.

Table 4 Summary of Sub-Scenarios in Scenario IX

Sub-Scenario	CMF for LW	CMF for SW	AF
IX-1	0.8	0.85	0.80
IX-2	0.8	0.85	0.90
IX-3	0.8	0.85	0.95
IX-4	0.8	0.85	1.05
IX-5	0.8	0.85	1.10
IX-6	0.8	0.85	1.20
IX-7	0.8	0.9	0.80
IX-8	0.8	0.9	0.90
IX-9	0.8	0.9	0.95
IX-10	0.8	0.9	1.05
IX-11	0.8	0.9	1.10
IX-12	0.8	0.9	1.20
IX-13	0.9	0.85	0.80
IX-14	0.9	0.85	0.90
IX-15	0.9	0.85	0.95
IX-16	0.9	0.85	1.05
IX-17	0.9	0.85	1.10
IX-18	0.9	0.85	1.20
IX-19	0.9	0.9	0.80
IX-20	0.9	0.9	0.90
IX-21	0.9	0.9	0.95
IX-22	0.9	0.9	1.05
IX-23	0.9	0.9	1.10
IX-24	0.9	0.9	1.20

Note: LW = lane width; SW = shoulder width.

The two coefficients for lane width and shoulder width (i.e., β_2 and β_3 in Equation 3-17) were used to estimate the CMFs for the two variables, respectively. The same indexes (i.e., estimation bias and error percentage) were used to evaluate the quality of CMFs derived from regression models, and the same GOF and prediction measures for the models were used for the regression models.

3.6 Data Description

This section describes the characteristics of the datasets used for the simulation analyses. Section 3.6.1 documents the datasets of roadway segments for Scenarios I to VI. Section 3.6.2 briefly summarizes the datasets for Scenarios VII and VIII (i.e., variable correlation.) And the datasets used in Scenario IX (i.e., combined safety effects of multiple treatments) is provided in Sections 3.6.3.

3.6.1 Datasets for Scenarios I to VI

The roadway data used in Scenarios I to VI contained 1,492 rural two-lane highway segments in Texas. The variables included segment length, AADT, lane width, curve density (i.e., curves/mile) and pavement friction. Pavement friction is the force that resists the relative motion between a vehicle tire and a pavement surface (Hall et al. 2009). Generally, higher pavement friction is linked to safer roads. The segment length and AADT were based on actual values from the Texas data, while the other three were hypothetical variables created specifically for this study. The lane widths were generated from a discrete uniform distribution with parameters 8 and 13. The curve density and pavement friction were generated from continuous uniform distributions. For the curve

density, the parameters were 0 and 16. And for pavement friction, the parameters were 16 and 48. The summary statistics of these variables are shown in Table 5.

Table 5 Summary Statistics of Highway Segments for Scenarios I to VI

Variable	Sample Size	Min.	Max	Mean (SD [c])
Length (mile)	1492	0.1	6.3	0.55 (0.67)
AADT	1492	502	24800	6643.9 (3996.4)
Lane Width (ft)	1492	8.0	13.0	10.47 (1.74)
CD [a] (per mile)	1492	0.02	16.0	8.1 (4.66)
PF [b]	1492	16.0	47.9	31.9 (9.08)

Note: a – CD = curve density; b – PF = pavement friction; c - SD = standard deviation.

3.6.2 Datasets for Scenarios VII and VIII

The considered variables in the previous section were generated independently, which might not be able to reflect the reality. Some highway characteristics are usually correlated with each other in practice. For example, roadways with higher AADT are more likely to be designed with wider lanes. In order to reflect the correlations between variables, the lane widths in this section were generated base on a multinomial logistic (MNL) regression analysis between AADT and lane width. The MNL model was developed from a real dataset, and it is shown in Table 6.

Table 6 MNL for Generating Lane Width (Baseline: 8 ft)

LW (ft)	Intercept	AADT (in 100,000)
9	2.202	-13.28
10	4.751	3.57
11	6.836	3.00
12	7.072	6.94
13	4.687	6.36

Note: LW = lane width.

So, for a given segment with AADT equal to *v* (in 100,000), the probabilities for the segment of having an 8-ft, 9-ft, 10-ft, 11-ft, 12-ft, or 13-ft lane, respectively, can be calculated as below:

$$P(LW=8 \mid AADT=v\times 10^5)=\frac{1}{1+e^{2.20-13.28*v}+e^{4.71+3.57*v}+\cdots+e^{4.69+6.36*v}} \quad (3\text{-}18a)$$

$$P(LW=9 \mid AADT=v\times 10^5)=\frac{e^{2.20-13.28*v}}{1+e^{2.20-13.28*v}+e^{4.71+3.57*v}+\cdots+e^{4.69+6.36*v}} \quad (3\text{-}18b)$$

$$P(LW=10 \mid AADT=v\times 10^5)=\frac{e^{4.71+3.57*v}}{1+e^{2.20-13.28*v}+e^{4.71+3.57*v}+\cdots+e^{4.69+6.36*v}} \quad (3\text{-}18c)$$

$$P(LW=11 \mid AADT=v\times 10^5)=\frac{e^{6.84+3.00*v}}{1+e^{2.20-13.28*v}+e^{4.71+3.57*v}+\cdots+e^{4.69+6.36*v}} \quad (3\text{-}18d)$$

$$P(LW=12 \mid AADT=v\times 10^5)=\frac{e^{7.07+6.94*v}}{1+e^{2.20-13.28*v}+e^{4.71+3.57*v}+\cdots+e^{4.69+6.36*v}} \quad (3\text{-}18e)$$

$$P(LW=13 \mid AADT=v\times 10^5)=\frac{e^{4.69+6.36*v}}{1+e^{2.20-13.28*v}+e^{4.71+3.57*v}+\cdots+e^{4.69+6.36*v}} \quad (3\text{-}18f)$$

Specifically, for a segment with 6,000 AADT, the six probabilities are 0.0003 (8 ft), 0.0013 (9 ft), 0.0447 (10 ft), 0.3474 (11 ft), 0.5568 (12 ft), and 0.0496 (13 ft), respectively.

Lane widths were generated for the 1,492 segments. Segment length and AADT were the same as those in Section 3.6.1. The summary statistics of these segments are shown in Table 7.

Table 7 Summary Statistics of Highway Segments for Scenarios VII and VIII

Variable	Sample Size	Min.	Max	Mean (SD)
Length (mile)	1,492	0.1	6.3	0.55 (0.67)
AADT	1,492	502	24800	6643.9 (3996.4)
Lane Width (ft)	1,492	8.0	13.0	11.62 (0.68)

Note: SD = standard deviation.

3.6.3 Datasets for Scenario IX

Scenario IX utilized the same roadway segments as those in the previous scenarios. The segment length and AADT were observed real data, while the two variables, lane width and shoulder width, were generated from discrete uniform distributions, respectively. Table 8 provides the summary statistics of the highway segments used in this scenario. Since both lane width and shoulder width had a discrete uniform distribution with two numbers, and they were independently generated, the four types of segment groups were equally distributed among all the segments. Each accounted for approximately 25%.

Table 8 Summary Statistics of Highway Segments for Scenario IX

Variable	Sample Size	Min.	Max	Mean (SD)
Length (mile)	1492	0.1	6.3	0.55 (0.67)
AADT	1492	502	24800	6643.9 (3996.4)
Lane Width (ft)	1492	12	13	12.5 (0.50)
Shoulder Width (ft)	1492	6	7	6.5 (0.50)

Note: SD = standard deviation.

It is important to point out that this study selected four geometric features and the CMFs were mainly assumed based on their practical values (i.e., from *HSM*, CMF Clearinghouse, etc.) to reflect as close as possible the characteristics related to variables that can influence crash risk. However, it does not have to be so. With the simulation protocol, it would be possible for other researchers to use variables and ranges based on characteristics associated with the roadway entities in which the researchers have detailed information on these characteristics.

3.7 Summary

This Chapter provided the methodologies regarding how to accomplish the four tasks. There are in total nine scenarios, as summarized below.

Scenario I: Consider one variable only, linear relationship.

Scenario II: Consider three variables, linear relationship.

Scenario III: Omitted variables, linear relationship.

Scenario IV: Consider one variable only, in nonlinear form.

Scenario V: Consider three variables, only one in nonlinear form.

Scenario VI: Consider three variables, two in nonlinear form.

Scenario VII: Variable Correlation, linear relationship.

Scenario VIII: Variable Correlation, nonlinear relationship.

Scenario IX: Combined safety effect.

The next chapter discusses the detailed evaluation of CMFs in each scenario.

4. ESTIMATING THE QUALITY OF CMFS*

Nine scenarios with numbers of sub-scenarios were analyzed using the methodologies described in Chapter 3. This chapter documents the results. Sections 4.1 to 4.5 provide the accuracy of CMFs for linear relationship, omitted variable, nonlinear relationship, variable correlation, and multiple treatments, respectively. The findings are summarized in Section 4.6.

4.1 Linear Relationships

This section documents the simulation results of Scenarios I and II, respectively.

4.1.1 Scenario I: Consider Lane Width Only

In this scenario, only the lane width was considered. All other factors affecting crash risk were assumed to be identical among all segments. This met the primary premise of cross-sectional studies that all locations were similar to each other in all other factors affecting crash risk.

The assumed CMF for lane width varied from 0.85 to 1.05 with an increment of 0.05. The theoretical function of the generated crash counts in this scenario is shown in Equation 4-1, which is similar to Equation 3-11, but the coefficient of lane width varied.

(4-1a)

* Part of this chapter is reprinted with permission from "Validation of Crash Modification Factors Derived from Cross-Sectional Studies with Regression Models" by L. Wu, D. Lord and Y. Zou, 2015. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2514, pp. 88-96, Washington, D.C. Copyright [2015] by the Transportation Research Board.

Or equivalently,

(4-1b)

Where,

β_{LW} = coefficient of lane width, varied between -0.163, -0.105, -0.051, 0 and 0.049, corresponding to the assumed CMFs equal to 0.85, 0.90, 0.95, 1.0 and 1.05, respectively; and,

β_{off} = offset coefficient, varied between 0.0019, 0.0010, 0.0005, 0.0003 and 0.0002, corresponding to the assumed CMFs equal to 0.85, 0.90, 0.95, 1.0 and 1.05, respectively.

The CMF for each assumed value was derived from the model using the same procedures illustrated in the simulation example (i.e., Section 3.1.2). The considered functional form is provided in Equation 3-12 (now 4-2 below).

β 2 (4-2)

The fitting results are shown in Table 9. It can be seen that for each of the assumed CMFs, the estimation bias between the CMF derived from the SPFs and the assumed value was relatively small under different simulation settings. The estimation bias was less than 0.005 for all scenarios, and the error was within 0.5 percent. The small standard deviation of CMFs (second column in Table 9) also indicated the CMFs derived from the experiments were consistent.

Table 9 Results of Scenario I

Theo. CMF [a]	CMF (SD) [b]	Bias	E [c]	AIC [d]	MAD [e]	MSPE [f]
ϕ = 0.5						
0.85	0.849 (0.014)	0.001	0.127	11127.63	0.047	0.013
0.90	0.901 (0.014)	-0.001	0.107	10681.33	0.042	0.011
0.95	0.949 (0.015)	0.001	0.114	10277.12	0.041	0.010
1.00	0.998 (0.015)	0.002	0.170	9901.71	0.037	0.008
1.05	1.051 (0.018)	-0.001	0.109	9540.71	0.036	0.008
ϕ = 1.0						
0.85	0.849 (0.018)	0.001	0.149	11289.108	0.063	0.026
0.90	0.902 (0.017)	-0.002	0.246	10800.501	0.053	0.017
0.95	0.945 (0.017)	0.005	0.498	10390.874	0.050	0.015
1.00	1.002 (0.024)	-0.002	0.233	9995.650	0.048	0.013
1.05	1.051 (0.019)	-0.001	0.076	9643.008	0.043	0.012

Table 9 Continued

Theo. CMF [a]	CMF (SD) [b]	Bias	E [c]	AIC [d]	MAD [e]	MSPE [f]
ϕ = 2.0						
0.85	0.853 (0.022)	-0.003	0.395	11020.611	0.083	0.044
0.90	0.899 (0.023)	0.001	0.099	10574.470	0.068	0.029
0.95	0.95 (0.024)	0.000	0.018 [g]	10185.386	0.065	0.026
1.00	0.996 (0.026)	0.004	0.436	9827.826	0.062	0.025
1.05	1.05 (0.025)	0.000	0.032 [g]	9454.310	0.056	0.022

Note: a – theoretical CMF; b – mean of CMFs from 100 experiments, SD is the standard deviation of the 100 CMFs; c – E is the error percentage, %; d, e, f – each is the mean value of the corresponding GOF measure of the 100 results; g – the non-zero error percentage with zero bias is caused by the rounding off during calculation.

Based on the result of Scenario I, it can be concluded that the CMFs derived from regression models can reflect the true safety performance of lane width when considering this variable only. In other words, if a regression model was based on a group of roadway segments that were ideally identical in all factors affecting traffic safety, except the segment length, AADT and lane width, the CMFs for lane width derived from the SPF should be unbiased.

4.1.2 Scenario II: Consider Three Variables with Fixed CMFs

In Scenario I, only one variable, the lane width, was considered. Scenario II considered a more practical condition: considering lane width, curve density and pavement friction. In this scenario, each of these three variables was assumed to have influence on crash risk, but they were not identical among all segments. This also met the primary premise of cross-sectional studies.

CMFs for lane width, curve density and pavement friction were assumed to be fixed, and they were 0.90, 1.072 and 0.973, respectively. For curve density, the 1.072 CMF meant that if the curve density of a segment increased by 1 per mile, the expected crash number would increase by 7.2 percent (1.072 - 1.0). And the baseline for curve density was 0 per mile. So, if the curve density of a segment was 0, the specific CMF for curve density of this segment was 1.0. For the pavement friction, the 0.973 CMF meant that if the pavement friction of a segment increased by 1 unit, the expected crash number would decrease by 2.7 percent (1.0 – 0.973). The baseline was 32.

The theoretical function form of the crash data in scenario II is shown in Equation 4-3. And the fitting equation for this scenario is shown in Equation 4-4.

$$N_{true,i} = N_{spf,i} \times CMF_{LW,i} \times CMF_{CD,i} \times CMF_{PF,i}$$

$$= 2.67 \times 10^{-4} \times L_i \times AADT_i \times 0.9^{LW_i - 12} \times 1.072^{CD_i} \times 0.973^{PF_i - 32} \quad \text{(4-3a)}$$

Or equivalently,

$$N_{true,i} = 0.0023 \times L_i \times AADT_i \times exp(-0.105 LW_i + 0.070 CD_i - 0.027 PF_i) \quad \text{(4-3b)}$$

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\beta_2 \times LW_i + \beta_3 \times CD_i + \beta_4 \times PF_i) \quad \text{(4-4)}$$

Where,

CD_i = curve density of segment i (per mile);

$CMF_{CD,i}$ = specific CMF value for curve density of segment i;

PF_i = pavement friction of segment i;

$CMF_{PF,i}$ = specific CMF value for pavement friction of segment i; and,

$\beta_2, \beta_3, \beta_4$ = coefficients to be estimated for lane width, curve density and pavement friction, respectively.

The result of this scenario is shown in Table 10. It can be seen that the CMFs derived from SPFs for all of the three variables were very close to the assumed values. The bias and error percentage were small. The result was quite similar as that of Scenario I. This means that, for fixed CMFs in this scenario, the regression model was able to derive reliable CMFs for the three variables. Furthermore, two other scenarios with more variables (five and eight in total, respectively) were analyzed, the results (not documented in this dissertation) were consistent with the results documented here with three variables.

Table 10 Results of Scenario II

Variable*	Theo .CMF[a]	CMF (SD) [b]	Bias	E [c]	AIC [d]	MAD[e]	MSPE[f]
ϕ = 0.5							
LW	0.900	0.900 (0.013)	0.000	0.023 [g]	13864.7	0.10	0.06
CD	1.072	1.073 (0.006)	-0.001	0.051			
PF	0.973	0.973 (0.002)	0.000	0.040 [g]			
ϕ = 1.0							
LW	0.900	0.897 (0.014)	0.003	0.366	14072.3	0.13	0.10
CD	1.072	1.072 (0.008)	0.000	0.019 [g]			
PF	0.973	0.973 (0.004)	0.000	0.026 [g]			
ϕ = 2.0							
LW	0.900	0.903 (0.023)	-0.003	0.354	13736.2	0.16	0.17
CD	1.072	1.072 (0.009)	0.000	0.032 [g]			
PF	0.973	0.972 (0.004)	0.001	0.063			

Note: a, b, c, d, e, f, g - the same notes as those in Table 9; * - LW = lane width, CD = curve density, PF = pavement friction.

Based on the results of this experiment, the CMFs derived from the commonly used GLMs can reflect the true safety performance when considering multiple variables and assuming other safety factors were identical among all segments.

4.2 Omitted Variables

This sections discusses Scenario III: consider three variables, but omit two in models. In Scenario II, although three variables were considered and included in the model, other factors affecting traffic safety were assumed to be identical among all segments. However, this was not the case in most crash prediction studies. Not all the factors affecting crashes were known or able to be captured by the model in practice. In this scenario, another condition was considered: both curve density and pavement friction were associated with crash risk, but only the lane width was included in the model.

The assumed CMFs for lane width, curve density and pavement friction were 0.90, 1.072 and 0.973, respectively, the same as those in Scenario II. The theoretical function of scenario III was the same as that of Scenario II, as shown in Equation 4-3b (now 4-5 below). In this scenario, the curve density and pavement friction were excluded from the model. So, the model for Scenario III was basically the same as that of Scenario I, as shown in Equation 3-12 or Equation 4-2 (now 4-6 below).

$$N_{true,i} = 0.0023 \times L_i \times AADT_i \times exp(-0.105 LW_i + 0.070 CD_i - 0.027 PF_i) \quad (4\text{-}5)$$

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\beta_2 \times LW_i) \quad (4\text{-}6)$$

Table 11 Results of Scenario III

Theo. CMF [a]	CMF (SD) [b]	Bias	E [c]	AIC [d]	MAD [e]	MSPE [f]
			ϕ = 0.5			
0.90	0.898 (0.014)	0.002	0.211	14395.9	0.75	2.45
			ϕ = 1.0			
0.90	0.890 (0.026)	0.010	1.111	14479.1	0.75	2.49
			ϕ = 2.0			
0.90	0.898 (0.023)	0.002	0.206	13975.8	0.75	2.51

Note: a, b, c, d, e, f - the same notes as those in Table 9.

The results for Scenario III are shown in Table 11. The derived CMF for lane width in this scenario was also close to the assumed value 0.90. The bias was relatively small and the error was within 1.2 percent in this experiment. Generally, the CMF for lane width in this experiment was reliable.

However, when compared with the results in Scenarios I and II, both the bias and error percentage became large in Scenario III. That is, when some factors affecting traffic safety were omitted in the models, the bias for the CMF might become higher. Meanwhile, the MAD and MSPE also increased greatly, indicating the modeling result became less reliable. Similar scenarios with CMFs for lane width of 0.85 as well as 1.05 and constant CMFs for curve density and pavement friction were analyzed, and the results were consistent.

In Scenario III, the assumed CMFs for curve density and pavement friction were close to 1.0. This means the change of one unit of these two variables would have relatively small effect or association on crash risk. In other words, if minor factors were omitted in the SPFs, the result might still be acceptable. However, the bias might become unacceptable if the omitted factors had a strong relationship with crashes. Further analyses were conducted to examine this hypothesis. For example, the assumed CMF for curve density was augmented to 1.2 and 1.3, which led to a significantly increase in the error. At the same time, the MAD and MSPE values also increased significantly. Therefore, when major factors were omitted in the SPFs, the CMFs derived may became unreliable.

In regression analysis, Mallows C_p has been proposed for model selection. The statistic C_p can be used as a criteria to assess fits when models with different numbers of parameters are being compared (Kutner et al. 2005). C_p is calculated by Equation 4-7.

$$C_p = \frac{RSS(p)}{MSE(full)} - N + 2 \times p \quad (4\text{-}7)$$

Where,

p = number of parameters in the subset model (i.e., omitted-variable model in this study);

$RSS(p)$ = residual sum of squares for the subset model;

$MSE(full)$ = mean square error for the full model (i.e., the model containing all variables affecting safety in this study); and,

N = sample size.

The subset model (or omitted-variable model) is considered to be "good" if $C_p \le p$. Note that "good" in this context means the omitted-variable model is acceptable, or the model is not suffering from omitted-variable bias.

The Mallows $C_p's$ were calculated in this scenario. With initial CMFs (0.90 for LW, 1.072 for CD and 0.973 for PF), the mean and standard error of the 100 $C_p's$ were 36.1 and 161, respectively. The probability that the omitted-variable model was "good" (i.e., $C_p \le 3$) was about 0.40. In other words, about 60% of these models were not "good" (i.e., they suffered from omitted-variable bias). This probability decreased

significantly as the assumed CMF for curve density increased. When the CMF for curve density was 1.3, it was nearly zero (i.e., nearly all the models suffered from omitted-variable bias). That is to say, when significant variables are excluded, the models suffer from the omitted-variable bias (Lord and Mannering 2010).

4.3 Nonlinear Relationships

This sections describes the findings of Scenarios IV, V and VI.

4.3.1 Scenario IV: Consider One Variable Only, Nonlinear Form

In this scenario, three nonlinear CM-Functions were assumed for lane width. This way, there were three sub-scenarios, IV-1, IV-2 and IV-3. The first two CM-Functions for lane width were quadratic functions (in logarithm form), shown in Equations 4-8 and 4-9.

$$ln(CMF) = 0.1 \times LW^2 - 2.22 \times LW + 12.28 \quad (4\text{-}8)$$

$$ln(CMF) = 0.2 \times LW^2 - 4.22 \times LW + 21.88 \quad (4\text{-}9)$$

The third one was a combination of two piecewise quadratic functions. This nonlinear function, shown in Equation 4-10, was developed by Lee et al. (2015) based on real crash data. Note that, in Lee et al. (2015)'s study lanes narrower than 9 ft were considered to have the same CMF as a 9-ft lane. To keep the analyses consistent and make it easier, this study assumed that an 8-ft lane had a different CMF with a 9-ft lane, and it was directly calculated using Equation 4-10.

$$ln(CMF) = \begin{cases} -0.11 \times (LW - 12)^2 + 0.30 & LW <= 12 \\ -0.08 \times (LW - 12)^2 + 0.30 & LW >= 12 \end{cases} \quad (4\text{-}10)$$

Table 12 Assumed CM-Functions for Lane Width in Scenario IV

Sub-Scenario	$ln(CMF)$ [a]	Line [b]	Area [c]	AVD	Level [d]
IV-1	$0.1\times LW^2-2.22\times LW+12.28$	$-0.123\times LW+1.46$	0.802	0.160	Weak
IV-2	$0.2\times LW^2-4.22\times LW+21.88$	$-0.023\times LW+0.24$	1.603	0.321	Strong
IV-3	$-0.11\times(LW-12)^2+0.30 \quad LW\le 12$ $-0.08\times(LW-12)^2+0.30 \quad LW\ge 12$	$0.339\times LW-0.45$	0.886	0.177	Weak

Note: a - LW = lane width (ft); b - Line = the closest line to the curve; c - Area = the area between the curve and its closest line; d - Level = the relative nonlinear level.

The assumed CM-Functions and their characteristics (closest line, area and AVD) for the three sub-scenarios are summarized in Table 12. It can be seen that, the AVD of Sub-Scenario IV-2 was higher than those of IV-1 and IV-3. The latter two were close to each other. This made the assumed CM-Function in IV-2 relatively strong in nonlinearity, and the other two relatively weak.

The theoretical function of the generated crash counts in these three sub-scenarios is shown in Equation 4-11. The specific CMF for lane width was calculated using Equations 4-8 to 4-10.

$$N_{true,i} = N_{spf,i} \times CMF_{LW,i} = 2.67 \times 10^{-4} \times L_i \times AADT_i \times CMF_{LW,i} \quad (4\text{-}11)$$

The considered functional form used in regression models is shown in Equation 4-12.

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\beta_2 \times LW_i) \quad (4\text{-}12)$$

Table 13 presents the CMFs derived from SPFs as well as other results (i.e., ϕ and GOF measurements) in this scenario. First, MAD and MSPE of nonlinear forms were significantly higher when compared with linear ones (i.e., Scenario I). This indicated the CMFs in this scenario might have higher bias. Second, with the increase of nonlinearity, the MAD and MSPE also increased. In other words, when the relationship between the variable and crash risk became strong in nonlinear level, the normal GLMs were likely to produce biased CMFs. Finally, under nonlinear relationships, in general, the inverse dispersion parameters estimated from SPFs were biased (see the column of "ϕ" in Table 13).

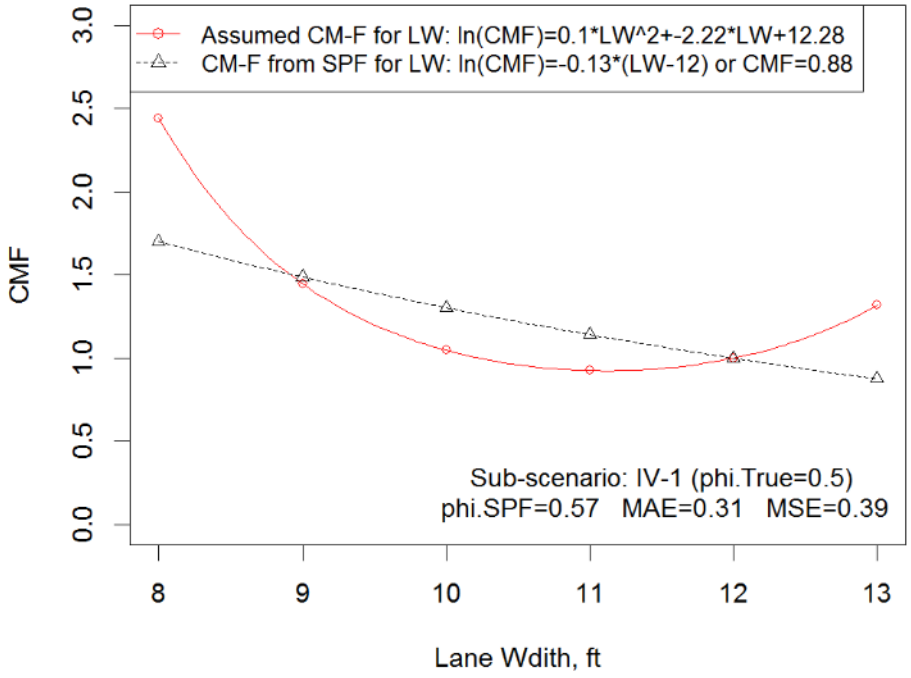
Table 13 Results of Scenario IV

# *	CMF (SD) [a]	ϕ [b]	AIC [d]	MAD [e]	MSPE [f]
ϕ [c] = 0.5					
IV-1	0.88 (0.01)	0.57	11413.39	0.31	0.39
IV-2	0.98 (0.02)	0.81	11522.30	0.64	1.57
IV-3	1.33 (0.02)	0.56	7770.30	0.15	0.11
ϕ [c] = 1.0					
IV-1	0.87 (0.02)	1.08	11492.77	0.31	0.39
IV-2	0.98 (0.02)	1.34	11509.57	0.64	1.58
IV-3	1.34 (0.03)	1.08	7803.04	0.15	0.13
ϕ [c] = 2.0					
IV-1	0.88 (0.03)	2.11	11233.07	0.32	0.40
IV-2	0.98 (0.03)	2.41	11138.83	0.64	1.60
IV-3	1.35 (0.03)	2.11	7690.76	0.15	0.15

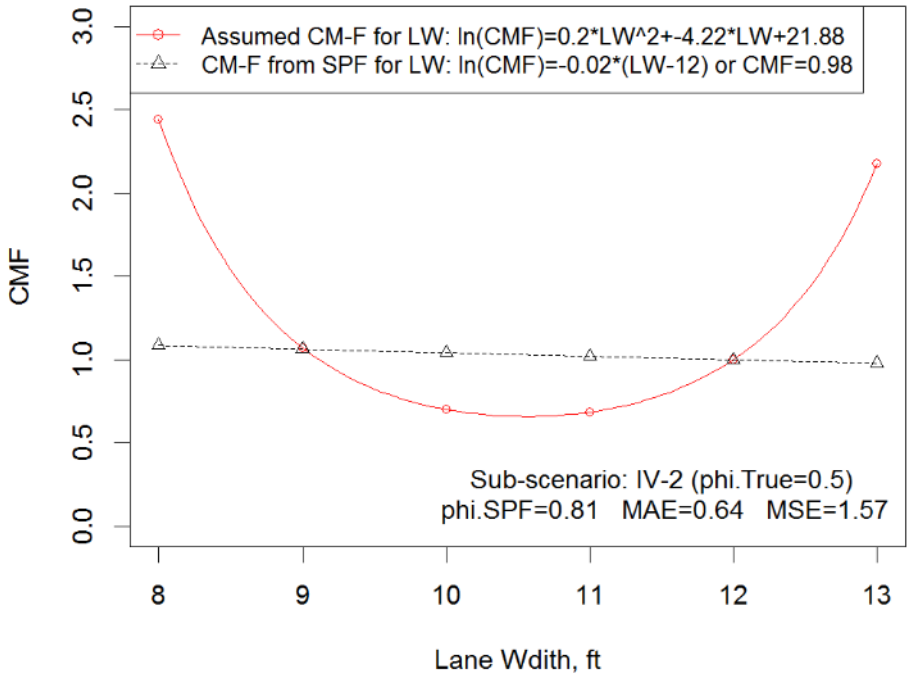
Note: a – mean of CMFs from 100 experiments, SD is the standard deviation of the 100 CMFs; b - the inverse dispersion parameter derived from SPFs; c – the theoretical inverse dispersion parameter in each sub-scenario; d, e, f – each is the mean value of the corresponding GOF measure of the 100 results; * - sub-scenario number.

To verify the above assumption, the curves of assumed CM-Functions and the CMFs derived from SPFs are illustrated in Figure 3. Figure 3 only presents the curves with a 0.5 inverse dispersion parameter. In addition, the specific CMFs for lane widths (8, 9, 11 and 13 ft) are presented in Table 14 for explicit comparison. The CMFs for 12-ft lane are excluded in Table 14, because 12 ft is the base condition for lane with and the CMFs are equal to 1.0 in both assumed and derived CM-Functions.

Figure 3(a) shows the CM-Functions in Sub-Scenario IV-1 (weak nonlinearity). It can be seen that the assumed true CMF for lane width first decreased from 8 ft until about 11 ft and then increased. But the CMF derived from SPF was 0.88, meaning the expected number of crashes consistently reduced by 12 percent whenever the lane was widened by 1 foot. According to the assumed true CMF, the 11-ft lane had the lowest crash risk. That is to say increasing lane width might bring negative influence on safety when the lane width was more than 11 ft. However, the CMF derived from regression analysis showed a contrary result, further safety benefits would be continually gained when widening lane to 12 or 13 ft. In addition, when the lane width was less than about 9 ft or more than 12 ft, the CMF was underestimated (the safety effect of widening lane was overestimated). The result was contrary when the lane width was between about 9 and 12 ft. The differences between the two were more obvious around boundary areas. The true CMF for an 8-ft lane was 2.44, whereas that derived from SPF was 1.78. The bias was 0.66 and error was 27.1 percent. Similar results can be observed when the lane was 13 ft. The specific CMFs, bias and error for other points (i.e., lanes with different widths) are shown in the rows of "IV-1" in Table 14.

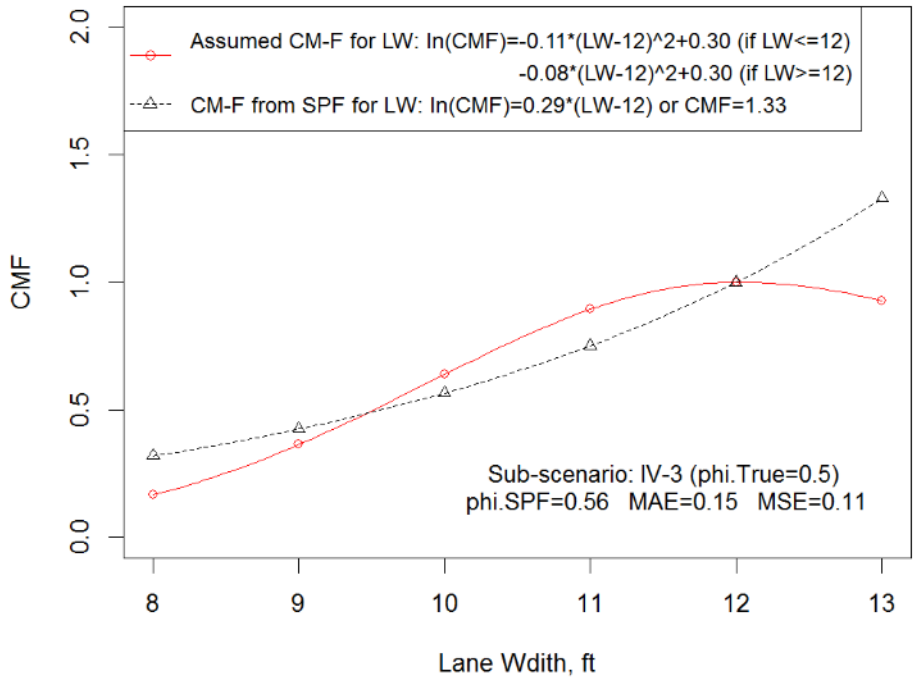


(a) Sub-Scenario IV-1



(b) Sub-Scenario IV-2

Figure 3 CM-Functions for Lane Width in Scenario IV (ϕ=0.5)



(c) Sub-Scenario IV-3

Figure 3 Continued

Table 14 Bias and Error of CMFs for Lane Width in Scenario IV

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
LW (ft)	8				9				10			
ϕ [d] = 0.5												
IV-1	2.44	1.78	0.66	27.1	1.45	1.54	-0.09	6.5	1.05	1.33	-0.29	27.4
IV-2	2.44	1.21	1.23	50.4	1.07	1.15	-0.08	7.7	0.70	1.10	-0.40	56.8
IV-3	0.17	0.31	-0.14	86.1	0.36	0.41	-0.05	13.7	0.64	0.56	0.08	13.0
ϕ [d] = 1.0												
IV-1	2.44	1.78	0.66	27.0	1.45	1.54	-0.09	6.6	1.05	1.33	-0.29	27.4
IV-2	2.44	1.19	1.25	51.2	1.07	1.14	-0.07	6.3	0.70	1.09	-0.39	55.4
IV-3	0.17	0.30	-0.13	80.0	0.36	0.40	-0.04	11.0	0.64	0.55	0.09	14.4
ϕ [d] = 2.0												
IV-1	2.44	1.73	0.71	28.9	1.45	1.51	-0.06	4.5	1.05	1.32	-0.27	25.7
IV-2	2.44	1.18	1.26	51.6	1.07	1.13	-0.06	5.7	0.70	1.09	-0.38	54.8
IV-3	0.17	0.28	-0.12	71.4	0.36	0.39	-0.03	7.0	0.64	0.53	0.10	16.5

Table 14 Continued

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
LW (ft)		11				13		
				ϕ [d]= 0.5				
IV-1	0.93	1.16	-0.23	24.7	1.32	0.87	0.45	34.4
IV-2	0.69	1.05	-0.36	52.9	2.17	0.95	1.22	56.2
IV-3	0.89	0.75	0.15	16.6	0.93	1.34	-0.41	44.7
				ϕ [d]= 1.0				
IV-1	0.93	1.16	-0.23	24.8	1.32	0.87	0.45	34.4
IV-2	0.69	1.04	-0.36	52.3	2.17	0.96	1.22	56.0
IV-3	0.89	0.74	0.15	17.3	0.93	1.35	-0.43	45.9
				ϕ [d]= 2.0				
IV-1	0.93	1.15	-0.22	23.9	1.32	0.87	0.45	33.9
IV-2	0.69	1.04	-0.36	51.9	2.17	0.96	1.22	55.9
IV-3	0.89	0.73	0.16	18.3	0.93	1.37	-0.44	47.7

Note: a – theoretical CMF (assumed true specific CMFs for lane widths of 8, 9, 10, 11 and 12 ft); b – CMF derived from SPF (specific CMFs derived from regression models for corresponding lane widths); c – error percentage, %; d – the theoretical inverse dispersion parameter (ϕ) in each sub-scenario; * - sub-scenario number.

Figure 3(b) presents the CM-Functions in Sub-Scenario IV-2 (strong nonlinearity). The overall result was similar to that of Sub-Scenario IV-1. CMFs derived from SPFs overestimated the safety effectiveness of lane width when it was less than 9 ft or more than 12 ft, and vice versa when it was between 9 and 12 ft. But the bias and error around the boundary areas in this sub-scenario were much higher than those in Sub-Scenario IV-1. For example, in this sub-scenario the bias at 8-ft lane was 1.23 and error was 50.4 percent, which were almost two times of those in Sub-Scenario IV-1. Bias and error for other points are also shown in Table 14 (rows of "IV-2"). More interestingly, the CMF derived from SPFs in this sub-scenario was 0.98, very close to 1.0, indicating lane width had minor influence on crash risk. Increasing lane width by one foot would only decrease crashes by about two percent. Safety analysts may misleadingly conclude that widening lane has little effect on reducing collisions based on this finding. However, the assumed true safety effect of lane width was far from this statement. In fact, both widening the lane from 11 ft and narrowing from10 ft would increase crash risk significantly.

The results of Sub-Scenario IV-3 (piecewise nonlinear functions) are shown in Figure 3(c). The CMF for lane width derived from SPFs was 1.33, widening the lane by one foot would increase crashes by 33 percent (1.33 – 1.0). When the lane width was between 9 and 12 ft, the two curves were close to each other. The error at 9-, 10- and 11-ft lanes were 13.7, 13.0 and 16.6 percentage, respectively. However, the bias was significantly high when the lane became relatively wide or narrow. The error reached nearly 90 percent at the point of 8-ft lane. On the side of wider lanes, the true CMF

decreased as lane width increased, but the CMF derived from SPFs increased continuously. Another interesting finding is that smaller MAD and/or MSPE did not always indicate smaller error percentage. When compared with the results in the other two sub-scenarios (i.e., Sub-Scenarios IV-1 and IV-2), the MAD and MSPE were consistently smaller in this sub-scenario. But the error percentages were significantly higher than those in the other two when the lane widths were 8 and 9 ft. This was probably due to the relatively smaller values of assumed CMFs in this sub-scenario. The assumed specific CMF for 8-ft lane was 0.17 in this sub-scenario, and a small bias lad to a relatively large error percentage under such condition (recall the definition for error percentage).

Similar results were found for other inverse dispersion parameters (i.e., 1.0 and 2.0). So, it can be concluded that none of the CMFs derived from SPFs could reflect the true safety effects accurately. They were all biased, especially around boundary areas. Regression analysis with commonly used linear link functions could produce biased CMFs when the variable had nonlinear relationships on crash risk. With the increase of nonlinearity, the bias became significant. In addition, the misuse of linear link function also led to biased estimates for other parameters, which might play important roles in safety analyses. For example, the inverse dispersion parameter is important in calculating the weights in EB analyses (Hauer et al. 2002; Wu et al. 2014; Zou et al. 2015). As a result, biased dispersion parameters lead to biased EB estimates of crashes.

4.3.2 Scenario V: Consider Three Variables, Only One in Nonlinear Form

In this scenario, three variables (i.e., lane width, curve density and pavement friction) were considered. Lane width was assumed to have a nonlinear relationship with crash risk. The other two were assumed to have linear relationships (i.e., fixed values). The assumed CMFs for curve density and pavement friction were 1.072 and 0.973, respectively. Actually the settings in this sub-scenario were basically the same as those of Scenario II, expect the lane width was assumed to have nonlinear relationships with crash risk. The same three nonlinear CM-Functions for lane width were used. Three sub-scenarios, V-1, V-2 and V-3, were analyzed, shown in Table 15.

Table 15 Assumed CM-Functions in Scenario V

# *	$ln(CMF)$ **(Nonlinear Level)**	*CMF*	
	Lane Width	**Curve Density**	**Pavement Friction**
V-1	$0.1\times LW^2-2.22\times LW+12.28$ (W)	1.072^{CD}	$0.973^{(PF-32)}$
V-2	$0.2\times LW^2-4.22\times LW+21.88$ (S)	1.072^{CD}	$0.973^{(PF-32)}$
V-3	$-0.11\times(LW-12)^2+0.30 \quad LW\le 12$ $-0.08\times(LW-12)^2+0.30 \quad LW\ge 12$ (W)	1.072^{CD}	$0.973^{(PF-32)}$

Note: * # = sub-scenario number; LW = lane width (ft); CD = curve density (number of curves per mile); PF = pavement friction.

The nonlinear level of the assumed CM-Functions for lane width of each sub-scenario was the same as the corresponding one in Scenario IV. It was relatively strong in Sub-Scenario V-2, and weak in V-1 and V-3.

The theoretical function of the generated crash counts and the considered functional form in this scenario are shown in Equations 4-13 and 4-14, respectively.

$$N_{true,i} = N_{spf,i} \times CMF_{LW,i} \times CMF_{CD,i} \times CMF_{PF,i}$$

$$= 2.67 \times 10^{-4} \times L_i \times AADT_i \times CMF_{LW,i} \times CMF_{CD,i} \times CMF_{PF,i} \quad (4\text{-}13)$$

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\beta_2 \times LW_i + \beta_3 \times CD_i + \beta_4 \times PF_i) \quad (4\text{-}14)$$

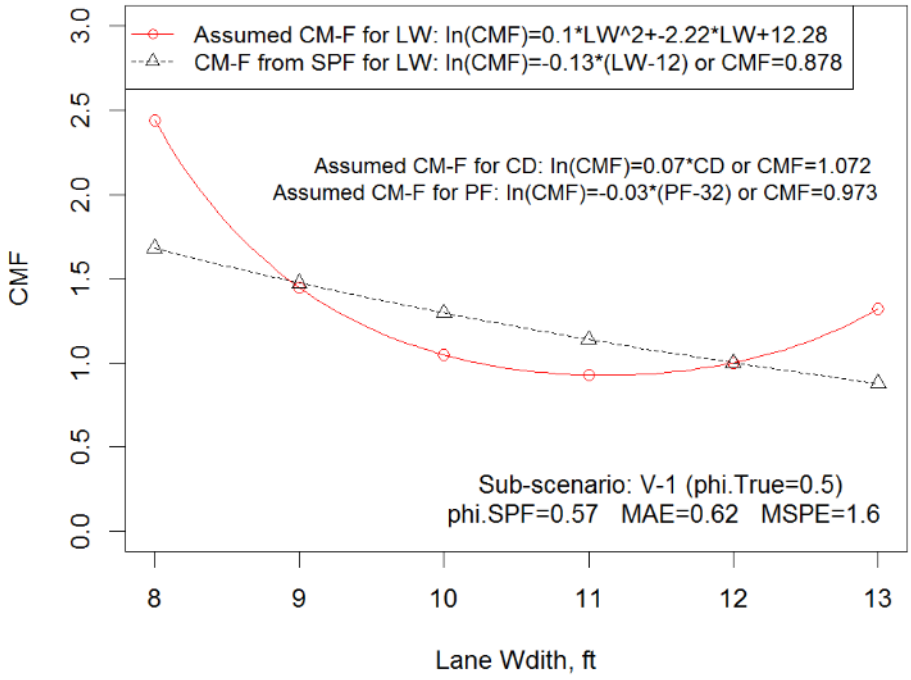
The CMFs for the three variables and other modeling results of each sub-scenario are documented in Table 16. The overall result was quite similar with that in Scenario IV. The MAD and MSPE were higher than those of linear relationships (i.e., Scenario II in the Section 4.1). Sub-Scenario V-2 consistently had the highest MAD and MSPE. Meanwhile, the inverse dispersion parameters estimated from SPFs were biased again. The second, third and fourth rows of Table 16 show the CMFs for lane width, curve density and pavement friction, respectively. The CMFs for lane width derived in this scenario were slightly different with those of Scenario IV, in which lane width was the only considered variable. The CMFs for curve density and pavement friction were very close to their true values. However, the MAD and MSPE of this scenario were higher than those of Scenario IV under the same assumed CM-Function for lane width.

Figure 4 illustrates the curves of assumed CM-Function for lane width and those derived from regression models in this scenario with a 0.5 inverse dispersion parameter. The specific CMFs for several lane widths of interest are provided in Table 17. The results were very close to those of the corresponding sub-scenario in Scenario IV. The CMFs were all biased, especially in boundary areas. The bias of Sub-Scenario V-2 was always higher than those of V-1 and V-3 (except over a very small range around 9).

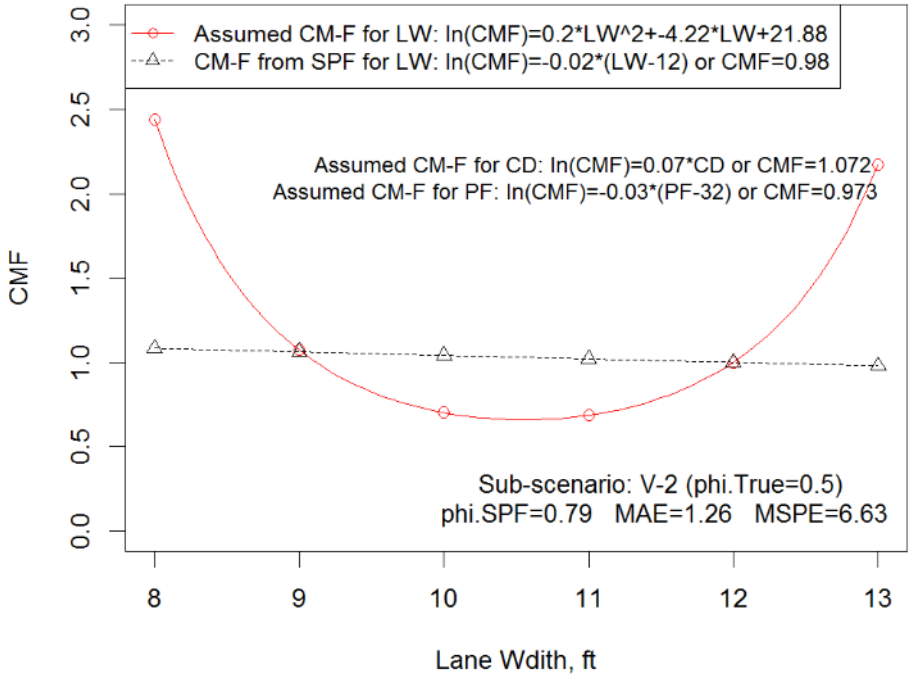
Table 16 Results of Scenario V

# *	CMF (SD) [a]			ϕ [b]	AIC [d]	MAD [e]	MSPE [f]
	LW	CD	PF				
ϕ [c] =0.5							
V-1	0.88 (0.014)	1.071 (0.007)	0.972 (0.003)	0.57	14677.2	0.62	1.60
V-2	0.98 (0.015)	1.069 (0.006)	0.971 (0.003)	0.79	14988.9	1.27	6.63
V-3	1.33 (0.021)	1.073 (0.006)	0.973 (0.003)	0.57	10441.8	0.29	0.52
ϕ [c] =1.0							
V-1	0.88 (0.015)	1.072 (0.007)	0.972 (0.003)	1.08	14842.6	0.62	1.65
V-2	0.98 (0.021)	1.069 (0.008)	0.971 (0.004)	1.31	14908.6	1.26	6.72
V-3	1.35 (0.024)	1.073 (0.008)	0.974 (0.004)	1.07	10520.1	0.30	0.60
ϕ [c] =2.0							
V-1	0.88 (0.022)	1.071 (0.008)	0.972 (0.004)	2.08	14439.9	0.63	1.76
V-2	0.98 (0.026)	1.069 (0.010)	0.97 (0.005)	2.36	14370.7	1.26	6.80
V-3	1.35 (0.034)	1.072 (0.010)	0.974 (0.005)	2.08	10277.6	0.31	0.66

Note: the same notes as those in Table 13.

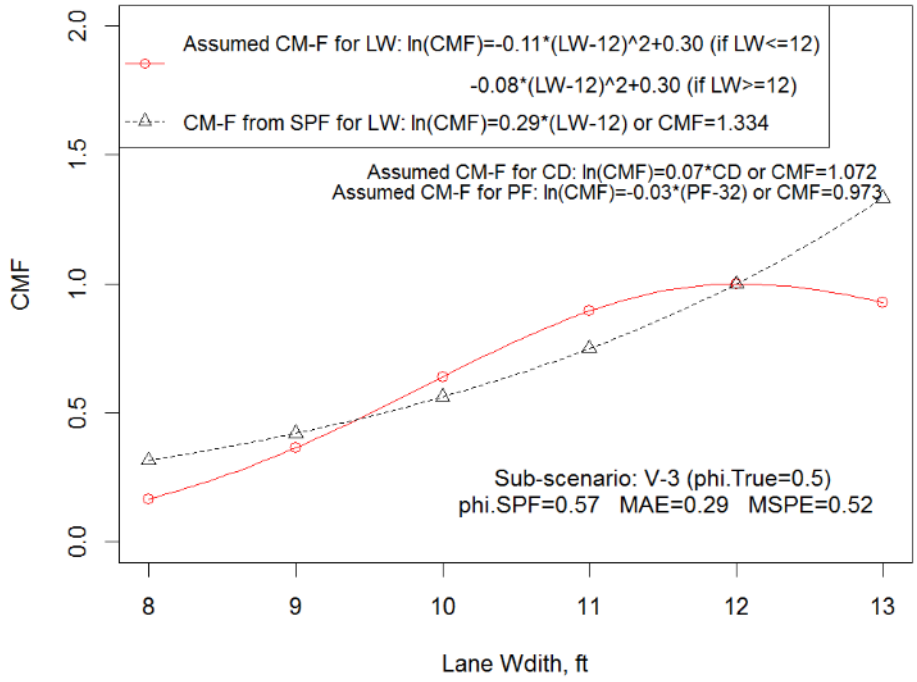


(a) Sub-Scenario V-1



(b) Sub-Scenario V-2

Figure 4 CM-Functions for Lane Width in Scenario V (ϕ=0.5)



(c) Sub-Scenario V-3

Figure 4 Continued

Table 17 Bias and Error of CMFs for Lane Width in Scenario V

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
LW (ft)	8				9				10			
ϕ [d] = 0.5												
V-1	2.44	1.68	0.76	31.2	1.45	1.48	-0.03	2.0	1.05	1.30	-0.25	23.7
V-2	2.44	1.08	1.36	55.6	1.07	1.06	0.01	0.9	0.70	1.04	-0.34	48.3
V-3	0.17	0.32	-0.15	90.3	0.36	0.42	-0.06	15.7	0.64	0.56	0.08	12.0
ϕ [d] = 1.0												
V-1	2.44	1.67	0.78	31.8	1.45	1.47	-0.02	1.3	1.05	1.29	-0.24	23.2
V-2	2.44	1.08	1.36	55.6	1.07	1.06	0.01	0.9	0.70	1.04	-0.34	48.3
V-3	0.17	0.30	-0.14	82.4	0.36	0.41	-0.04	12.0	0.64	0.55	0.09	13.8
ϕ [d] = 2.0												
V-1	2.44	1.64	0.80	32.7	1.45	1.45	0.00	0.3	1.05	1.28	-0.23	22.4
V-2	2.44	1.08	1.36	55.9	1.07	1.06	0.01	1.3	0.70	1.04	-0.34	47.9
V-3	0.17	0.30	-0.14	81.6	0.36	0.41	-0.04	11.7	0.64	0.55	0.09	14.0

Table 17 Continued

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
LW (ft)	11				13			
ϕ [d] = 0.5								
V-1	0.93	1.14	-0.21	22.9	1.32	0.88	0.44	33.4
V-2	0.69	1.02	-0.33	48.7	2.17	0.98	1.19	54.9
V-3	0.89	0.75	0.14	16.1	0.93	1.33	-0.41	43.9
ϕ [d] = 1.0								
V-1	0.93	1.14	-0.21	22.7	1.32	0.88	0.44	33.3
V-2	0.69	1.02	-0.33	48.8	2.17	0.98	1.19	54.9
V-3	0.89	0.74	0.15	17.0	0.93	1.35	-0.42	45.4
ϕ [d] = 2.0								
V-1	0.93	1.13	-0.21	22.3	1.32	0.88	0.44	33.0
V-2	0.69	1.02	-0.33	48.5	2.17	0.98	1.19	54.9
V-3	0.89	0.74	0.15	17.1	0.93	1.35	-0.42	45.6

Note: a, b, c, d, * – the same notes as those in Table 14; LW = lane width.

The CMFs for curve density derived from SPFs in the three sub-scenarios (with a 0.5 inverse dispersion parameter) were 1.071, 1.069, and 1.073, respectively. They were quite close to the assumed true value, 1.072. The curves for the four CM-Functions are shown in Figure 5. The specific CMFs, bias as well as error percentages at some points are listed in Table 18. The CMFs were generally acceptable. However, when comparing the results between the three sub-scenarios, it can be observed that the bias and error percentage in Sub-Scenario V-2 (strong nonlinearity) were always higher than those in V-1 and V-3 (weak nonlinearity). So, as the nonlinearity between lane width and crash risk increased, the bias of CMF for curve density became significant. That is to say, even the link function for one variable was correct, the accuracy of CMF for this variable can still be influenced if incorrect or improper link functions for other variables had been utilized in the models.

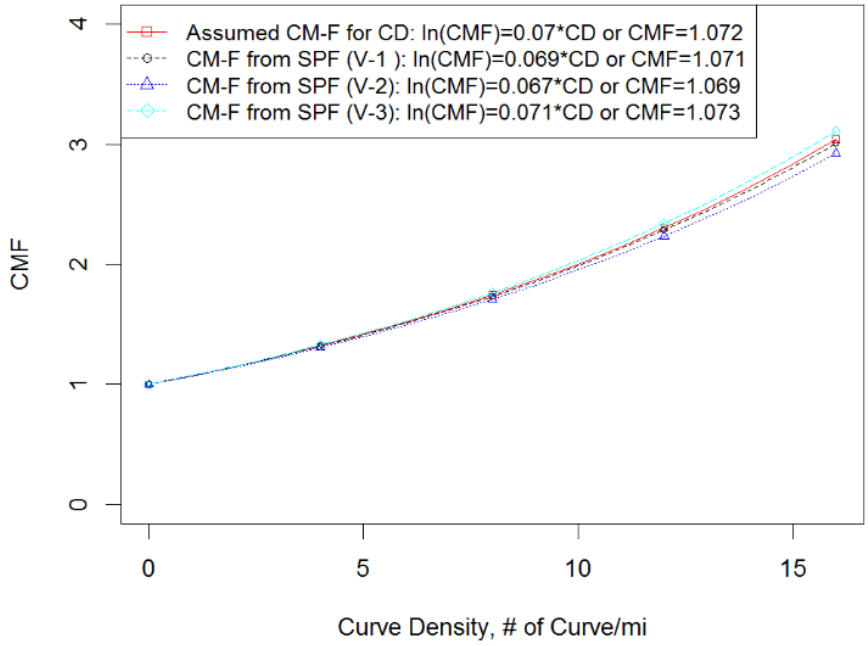


Figure 5 CM-Functions for Curve Density in Scenario V (ϕ=0.5)

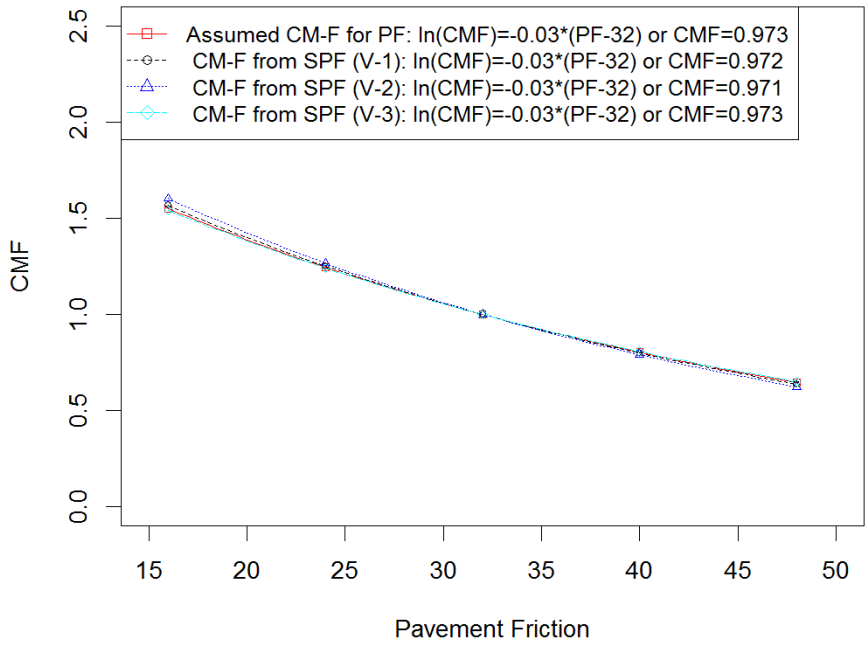


Figure 6 CM-Functions for Pavement Friction in Scenario V (ϕ=0.5)

Table 18 Bias and Error of CMFs for Curve Density in Scenario V

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
CD	4				8			
ϕ [d] = 0.5								
V-1	1.32	1.32	0.00	0.3	1.74	1.73	0.01	0.6
V-2	1.32	1.31	0.01	1.0	1.74	1.71	0.03	2.0
V-3	1.32	1.33	-0.01	0.5	1.74	1.76	-0.02	1.1
ϕ [d] = 1.0								
V-1	1.32	1.32	0.00	0.1	1.74	1.74	0.00	0.2
V-2	1.32	1.31	0.01	0.9	1.74	1.71	0.03	1.9
V-3	1.32	1.32	0.00	0.3	1.74	1.75	-0.01	0.5
ϕ [d] = 2.0								
V-1	1.32	1.32	0.00	0.4	1.74	1.73	0.01	0.7
V-2	1.32	1.31	0.01	1.0	1.74	1.71	0.03	2.0
V-3	1.32	1.32	0.00	0.0	1.74	1.74	0.00	0.1

Table 18 Continued

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
CD	12				16			
ϕ [d] = 0.5								
V-1	2.30	2.28	0.02	0.9	3.04	3.01	0.04	1.2
V-2	2.30	2.24	0.07	2.9	3.04	2.92	0.12	3.9
V-3	2.30	2.34	-0.04	1.7	3.04	3.11	-0.07	2.2
ϕ [d] = 1.0								
V-1	2.30	2.30	0.01	0.3	3.04	3.03	0.01	0.4
V-2	2.30	2.24	0.06	2.8	3.04	2.93	0.11	3.7
V-3	2.30	2.32	-0.02	0.8	3.04	3.07	-0.03	1.1
ϕ [d] = 2.0								
V-1	2.30	2.28	0.02	1.0	3.04	3.00	0.04	1.4
V-2	2.30	2.24	0.07	3.0	3.04	2.92	0.12	3.9
V-3	2.30	2.30	0.00	0.1	3.04	3.04	0.00	0.1

Note: a, b, c, d, * – the same notes as those in Table 14; CD = curve density (number of curves per mile).

Table 19 Bias and Error of CMFs for Pavement Friction in Scenario V

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
PF	16				24			
ϕ [d] = 0.5								
V-1	1.55	1.57	-0.02	1.2	1.24	1.25	-0.01	0.6
V-2	1.55	1.60	-0.05	3.5	1.24	1.27	-0.02	1.7
V-3	1.55	1.54	0.01	0.5	1.24	1.24	0.00	0.2
ϕ [d] = 1.0								
V-1	1.55	1.57	-0.02	1.1	1.24	1.25	-0.01	0.5
V-2	1.55	1.60	-0.05	3.5	1.24	1.27	-0.02	1.8
V-3	1.55	1.53	0.01	0.9	1.24	1.24	0.01	0.5
ϕ [d] = 2.0								
V-1	1.55	1.56	-0.02	1.0	1.24	1.25	-0.01	0.5
V-2	1.55	1.62	-0.07	4.4	1.24	1.27	-0.03	2.2
V-3	1.55	1.54	0.01	0.8	1.24	1.24	0.01	0.4

Table 19 Continued

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
PF	40				48			
ϕ [d] = 0.5								
V-1	0.80	0.80	0.00	0.6	0.65	0.64	0.01	1.2
V-2	0.80	0.79	0.01	1.7	0.65	0.62	0.02	3.4
V-3	0.80	0.81	0.00	0.3	0.65	0.65	0.00	0.5
ϕ [d] = 1.0								
V-1	0.80	0.80	0.00	0.5	0.65	0.64	0.01	1.0
V-2	0.80	0.79	0.01	1.7	0.65	0.62	0.02	3.4
V-3	0.80	0.81	0.00	0.5	0.65	0.65	-0.01	1.0
ϕ [d] = 2.0								
V-1	0.80	0.80	0.00	0.5	0.65	0.64	0.01	1.0
V-2	0.80	0.79	0.02	2.1	0.65	0.62	0.03	4.2
V-3	0.80	0.81	0.00	0.4	0.65	0.65	-0.01	0.9

Note: a, b, c d, * – the same notes as those in Table 14; PF = pavement friction.

The CMFs for pavement friction produced from the three sub-scenarios (with a 0.5 inverse dispersion parameter) were 0.972, 0.971 and 0.973, respectively. The CM-Function curves are shown in Figure 6. Specific values of CMFs as well as bias and error percentage at some points are listed in Table 19. The results were similar with those of curve density. Overall, the bias and error percentage were relatively small. And the highest error percentage appeared around boundary areas. The further the point was away from the baseline, the higher the bias and error percentage were. The error percentage was under 5 percent, indicating the results were acceptable in all the three sub-scenarios. However, Sub-Scenario V-2 was consistently the highest in terms of bias and error percentage. So, the CMFs for pavement friction derived from SPFs were likely to become less accurate as the nonlinear level of lane width became strong.

Another interesting finding worth to mention is the relationship between the quality of CMFs and the GOF of the model results (i.e., MAD and MSPE). The MAD and MSPE were higher in this scenario than those in Scenario IV (see the last two columns of Table 13 and Table 16). It seems that adding another two variables made the modeling result less accurate, which might potentially reduce the quality of CMFs derived from SPFs. So, one might assume the CMFs of this scenario should have higher bias and error than those of Scenario IV. However, it was not always true. Comparison between Table 14 and Table 17 indicates the CMFs for lane width in this scenario generally only had slightly higher error percentage than those of Scenario IV except a small range around 9 ft. One possible reason was that two additional variables were included in this scenario, and they both had considerable influence on the response

variable (i.e., true crash mean). They might have potentially influenced the quality of CMFs for all the three variables. Another possible reason was the differences of response variables (i.e., expected crash count) in the two scenarios. Simple comparison showed the mean of the random crash numbers in this scenario was about two times of that of Scenario IV. So it is not surprising the MAD and MSPE were higher in this scenario given other conditions (e.g., sample size, model link functions, etc.) were similar in the two.

Similar results were found when the inverse dispersion parameter was 1.0 and 2.0. So the main findings of this scenario can be summarized as follows: (1) the CM-Function for lane width derived from the common regression models (i.e., GLMs) were biased when it had a nonlinear relationship with crash risk and improper function form was used in the regression models; (2) with the increase of nonlinearity (i.e. nonlinear relationship became stronger), the bias trended to become more significant; (3) the CMFs for other variables having linear relationship might be acceptable when mixed with those having nonlinear relationship. But the quality decreased as the nonlinear relationship became stronger; (4) the misuse of linear link function for one or more variables also led to biased estimates of other parameters.

4.3.3 Scenario VI: Consider Three Variables, Two in Nonlinear Form

In this scenario, the three variables (i.e., lane width, curve density and pavement friction) were considered again. But two of them, lane width and curve density, were assumed to have nonlinear relationships with crash risk. Pavement friction was assumed to have a linear relationship (i.e., fixed value).

To simplify the analyses, the first two nonlinear CM-Functions for lane width in Scenarios IV and V were used in this scenario, and the last one with piecewise function was removed. The assumed CMF for pavement friction was 0.973, the same as that in Scenario V.

Two quadratic CM-Functions for curve density were assumed, as shown in Equations 4-15 and 4-16, respectively.

$$ln(CMF_i) = 8.7 \times 10^{-4} \times CD_i^2 + 5.56 \times 10^{-2} \times CD_i \quad (4\text{-}15)$$

$$ln(CMF_i) = 3.5 \times 10^{-3} \times CD_i^2 + 1.39 \times 10^{-2} \times CD_i \quad (4\text{-}16)$$

Where,

CMF_i = the specific CMF for curve density of segment *i*; and

CD_i = curve density of the segment, (average numbers of curve per mile).

The CM-Functions for curve density and other characteristics regarding nonlinear level (the closest line to the curve, area between the two, and average vertical distance) are listed in Table 20. It can be seen that both area and average vertical distance of the second function are much higher than those of the first one. So, the nonlinear level of the second one is stronger than the first one.

Table 20 Assumed CM-Functions for Curve Density in Scenario VI

$ln(CMF)$ [a]	Line [b]	Area[c]	AVD[d]	Level[e]
$8.7\times10^{-4}CD^2+5.56\times10^{-2}CD$	$6.95\times10^{-2}CD-0.037$	0.229	0.014	Weak
$3.5\times10^{-3}CD^2+1.39\times10^{-2}CD$	$6.99\times10^{-2}CD-0.149$	0.920	0.057	Strong

Note: a - LW = lane width (ft); b - Line = the closes line to the curve; c - Area = the area between the curve and its closest line; d - AVD = average vertical distance between the curve and the line; e - Level = the relative nonlinear level.

In total, there were four sub-scenarios in this scenario, shown in Table 21. It can be seen that the nonlinear level of Sub-Scenario VI-1 was weak in both lane width and curve density. That of Sub-Scenario VI-4 was strong in both. Sub-Scenarios VI-2 and VI-3 were a combination of a weak and a strong.

The theoretical function of the generated crash counts and considered functional form used in this scenario were identical with those in Scenario V (i.e., Equations 4-13 and 4-14). They are reproduced below as Equations 4-17 and 4-18, respectively.

Table 21 Assumed CM-Functions in Scenario VI

# *	$ln(CMF)$ **(Nonlinear Level)**		CMF
	Lane Width	**Curve Density**	**Pavement Friction**
VI-1	$0.1LW^2 - 2.22LW + 12.28$ (W)	$8.7\times10^{-4}CD^2 + 5.56\times10^{-2}CD$ (W)	$0.973^{(PF-32)}$
VI-2	$0.2LW^2 - 4.22LW + 21.88$ (S)	$8.7\times10^{-4}CD^2 + 5.56\times10^{-2}CD$ (W)	$0.973^{(PF-32)}$
VI-3	$0.1LW^2 - 2.22LW + 12.28$ (W)	$3.5\times10^{-3}CD^2 + 1.39\times10^{-2}CD$ (S)	$0.973^{(PF-32)}$
VI-4	$0.2LW^2 - 4.22LW + 21.88$ (S)	$3.5\times10^{-3}CD^2 + 1.39\times10^{-2}CD$ (S)	$0.973^{(PF-32)}$

Note: the same notes as those in Table 15.

$$N_{true,i} = N_{spf,i} \times CMF_{LW,i} \times CMF_{CD,i} \times CMF_{PF,i}$$

$$= 2.67 \times 10^{-4} \times L_i \times AADT_i \times CMF_{LW,i} \times CMF_{CD,i} \times CMF_{PF,i} \quad (4\text{-}17)$$

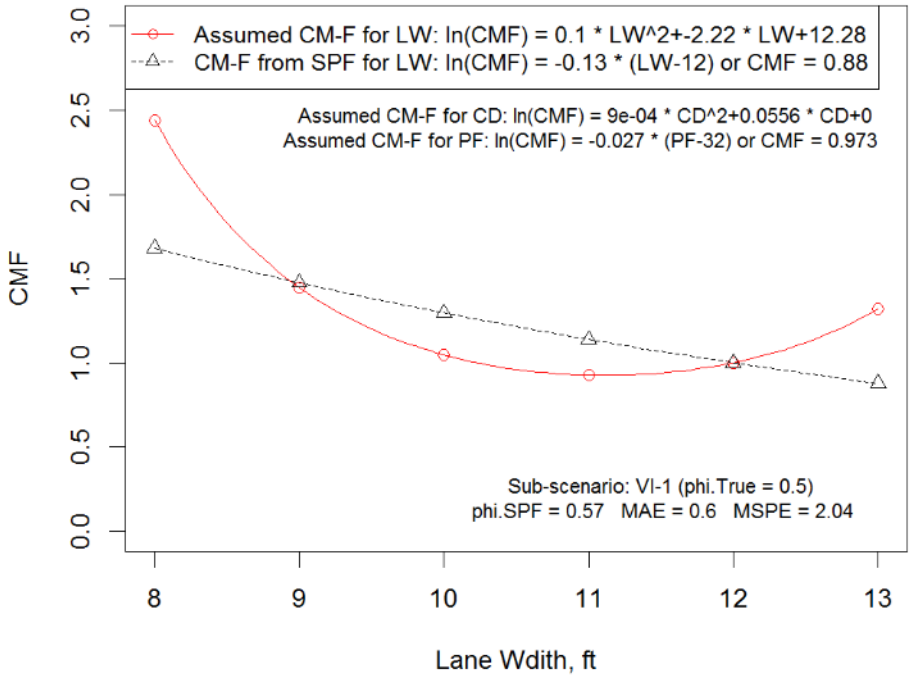
$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\beta_2 \times LW_i + \beta_3 \times CD_i + \beta_4 \times PF_i) \quad (4\text{-}18)$$

The CMFs for the three variables as well as other results produced from the modeling are presented in Table 22. Similarly to Scenario V, the MAD and MSPE were higher than those of linear relationships (i.e., Scenario II in Section 4.1). But surprisingly they were always the highest in Sub-Scenario VI-2 (combination of strong and weak) rather than in VI-4 (strong in both). The inverse dispersion parameters estimated from SPFs were biased again in this scenario. The second, third and fourth rows of Table 22 show the CMFs for lane width, curve density and pavement friction, respectively. The CMF for lane width derived in this scenario was nearly the same as that of Scenario V with corresponding assumed CM-Function. The CMFs for curve density were slightly different with those in Scenario V. And the CMFs for pavement friction were very close to the true value.

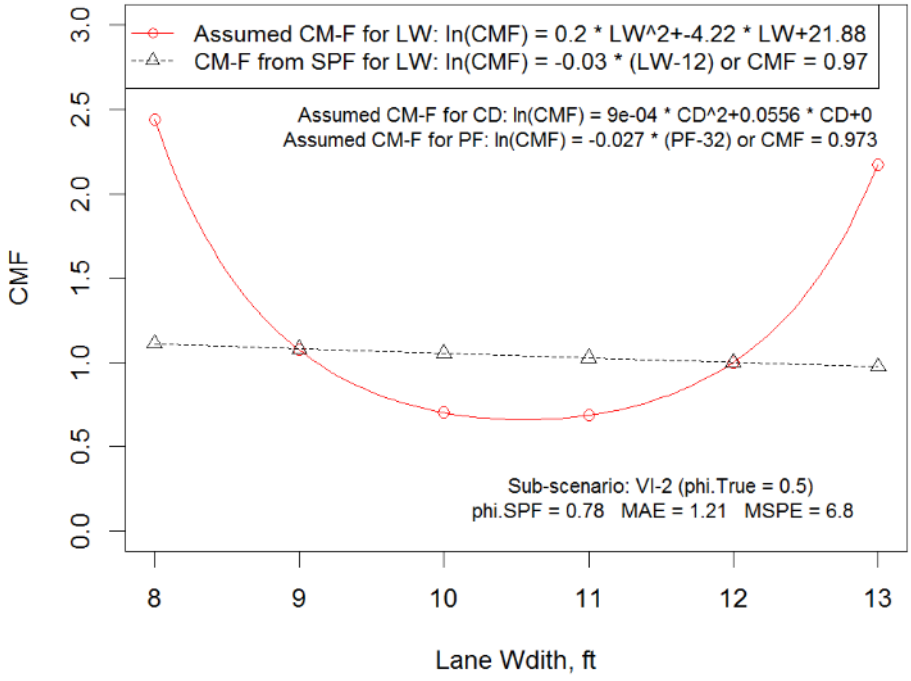
Table 22 Results of Scenario VI

# *	CMF (SD) [a]			ϕ [b]	AIC [d]	MAD [e]	MSPE [f]
	LW	CD	PF				
			ϕ [c]=0.5				
VI-1	0.88 (0.014)	1.073 (0.006)	0.973 (0.002)	0.57	14455.8	0.60	2.05
VI-2	0.97 (0.017)	1.072 (0.005)	0.972 (0.002)	0.79	14670.4	1.21	6.80
VI-3	0.88 (0.015)	1.075 (0.005)	0.973 (0.003)	0.57	13835.8	0.55	1.61
VI-4	0.97 (0.016)	1.074 (0.006)	0.972 (0.003)	0.79	13980.7	1.08	5.47
			ϕ [c]=1.0				
VI-1	0.88 (0.015)	1.072 (0.008)	0.972 (0.003)	1.07	14648.4	0.61	2.15
VI-2	0.97 (0.020)	1.074 (0.008)	0.972 (0.003)	1.31	14621.6	1.21	6.89
VI-3	0.88 (0.018)	1.073 (0.008)	0.973 (0.003)	1.08	13965.8	0.55	1.71
VI-4	0.97 (0.021)	1.074 (0.008)	0.972 (0.004)	1.32	13978.6	1.08	5.55
			ϕ [c]=2.0				
VI-1	0.88 (0.021)	1.073 (0.010)	0.972 (0.005)	2.09	14243.5	0.62	2.26
VI-2	0.98 (0.027)	1.073 (0.010)	0.971 (0.005)	2.36	14118.3	1.22	7.16
VI-3	0.88 (0.022)	1.073 (0.010)	0.973 (0.004)	2.09	13592.0	0.56	1.81
VI-4	0.98 (0.027)	1.074 (0.011)	0.971 (0.004)	2.36	13420.4	1.08	5.70

Note: the same notes as those in Table 13.

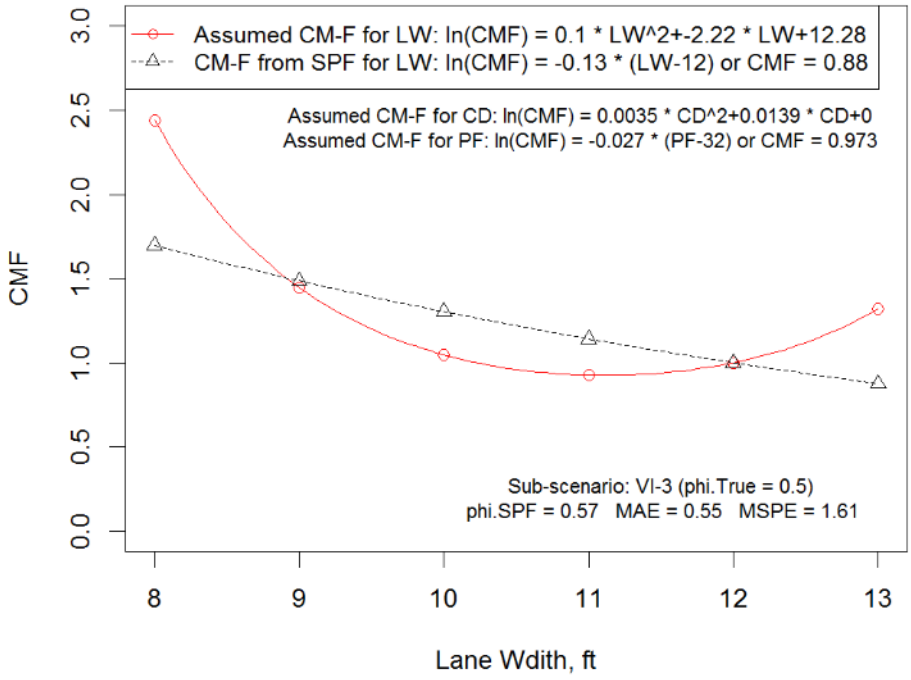


(a) Sub-Scenario VI-1

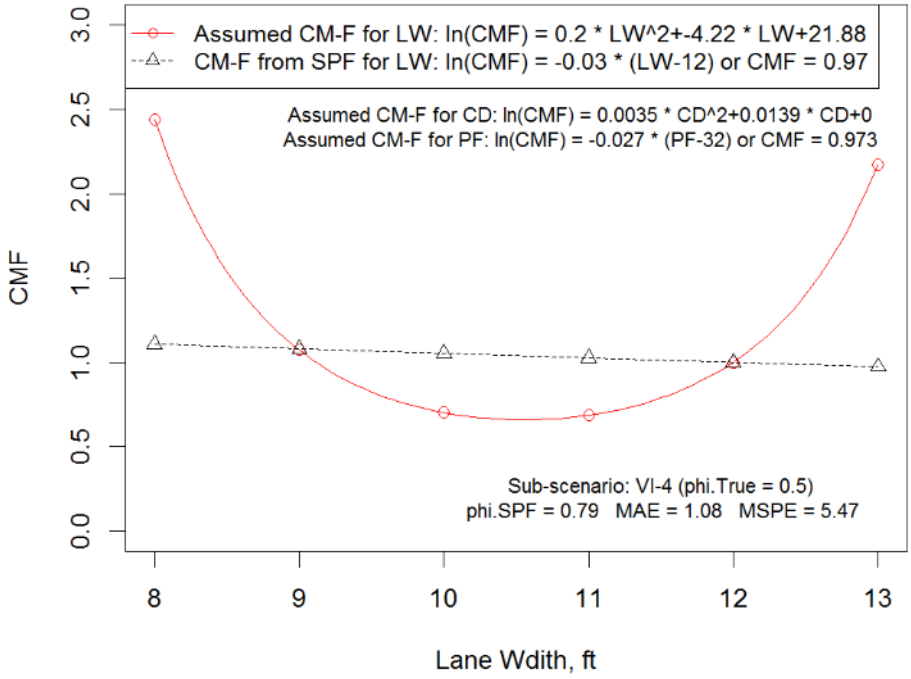


(b) Sub-Scenario VI-2

Figure 7 CM-Functions for Lane Width in Scenario VI (ϕ=0.5)



(c) Sub-Scenario VI-3



(d) Sub-Scenario VI-4

Figure 7 Continued

Table 23 Bias and Error of CMFs for Lane Width in Scenario VI

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
LW (ft)	8				9				10			
ϕ [d]= 0.5												
VI-1	2.44	1.66	0.78	31.91	1.45	1.46	-0.02	1.18	1.05	1.29	-0.24	23.10
VI-2	2.44	1.07	1.37	56.26	1.07	1.05	0.02	1.99	0.70	1.03	-0.33	47.19
VI-3	2.44	1.67	0.77	31.71	1.45	1.47	-0.02	1.41	1.05	1.29	-0.24	23.28
VI-4	2.44	1.06	1.38	56.42	1.07	1.05	0.02	2.27	0.70	1.03	-0.33	46.92
ϕ [d]= 1.0												
VI-1	2.44	1.65	0.79	32.55	1.45	1.45	-0.01	0.47	1.05	1.28	-0.24	22.52
VI-2	2.44	1.06	1.38	56.45	1.07	1.05	0.02	2.31	0.70	1.03	-0.33	46.87
VI-3	2.44	1.64	0.80	32.81	1.45	1.45	0.00	0.18	1.05	1.28	-0.23	22.28
VI-4	2.44	1.07	1.37	56.12	1.07	1.05	0.02	1.76	0.70	1.04	-0.33	47.43
ϕ [d]= 2.0												
VI-1	2.44	1.66	0.78	31.87	1.45	1.46	-0.02	1.23	1.05	1.29	-0.24	23.14
VI-2	2.44	1.08	1.36	55.76	1.07	1.06	0.01	1.15	0.70	1.04	-0.34	48.03
VI-3	2.44	1.63	0.81	33.32	1.45	1.44	0.01	0.39	1.05	1.28	-0.23	21.82
VI-4	2.44	1.07	1.37	56.32	1.07	1.05	0.02	2.09	0.70	1.03	-0.33	47.09

Table 23 Continued

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
LW (ft)	11				13			
				ϕ [d]= 0.5				
VI-1	0.93	1.14	-0.21	22.62	1.32	0.88	0.44	33.23
VI-2	0.69	1.02	-0.33	48.18	2.17	0.98	1.19	54.76
VI-3	0.93	1.14	-0.21	22.71	1.32	0.88	0.44	33.28
VI-4	0.69	1.02	-0.33	48.05	2.17	0.98	1.19	54.72
				ϕ [d]= 1.0				
VI-1	0.93	1.13	-0.21	22.33	1.32	0.88	0.44	33.07
VI-2	0.69	1.02	-0.33	48.02	2.17	0.98	1.19	54.72
VI-3	0.93	1.13	-0.21	22.21	1.32	0.88	0.44	33.01
VI-4	0.69	1.02	-0.33	48.30	2.17	0.98	1.19	54.80
				ϕ [d]= 2.0				
VI-1	0.93	1.14	-0.21	22.64	1.32	0.88	0.44	33.24
VI-2	0.69	1.02	-0.33	48.61	2.17	0.98	1.19	54.89
VI-3	0.93	1.13	-0.20	21.98	1.32	0.89	0.43	32.88
VI-4	0.69	1.02	-0.33	48.13	2.17	0.98	1.19	54.75

Note: a, b, c, d, * – the same notes as those in Table 14.

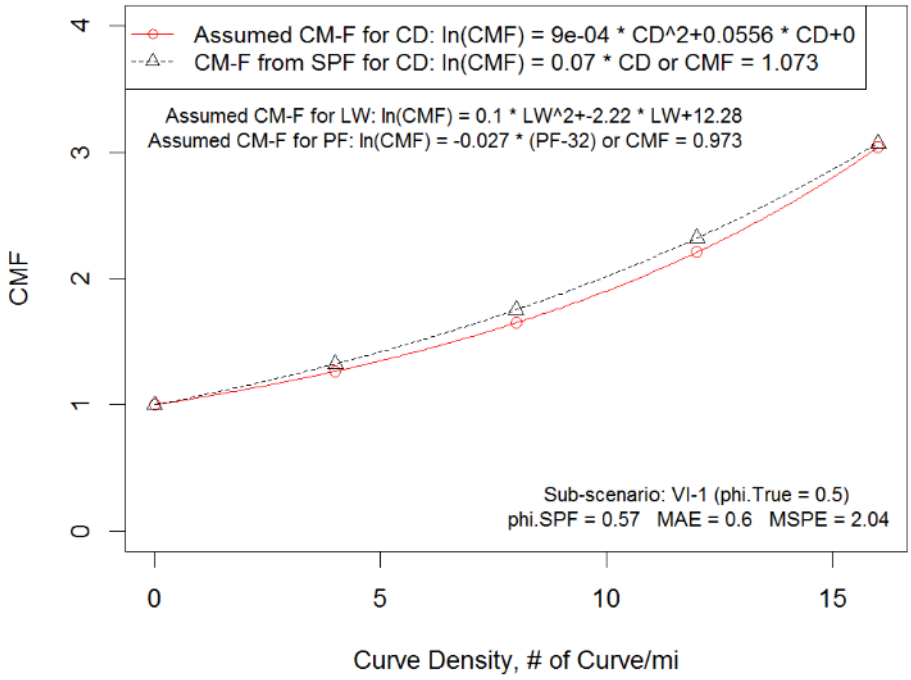
Figure 7 illustrates the curves of assumed CM-Functions for lane width and those derived from regression models in this scenario with a 0.5 inverse dispersion parameter. The specific CMFs for several lane widths of interest are provided in Table 23. The results of lane width were nearly identical with those of the corresponding one in Scenario V. The CMFs were all biased, especially around boundary areas. The calculation indicated the bias in Sub-Scenarios VI-1 and VI-3 (weak in lane width) were significantly lower than those in VI-2 and VI-4 (strong in lane width). It seems the changes in nonlinearity of curve density had no significant influence on the CMF for lane width.

The CMFs for curve density derived from SPFs in the four sub-scenarios with a 0.5 inverse dispersion parameter were 1.073, 1.072, 1.075 and 1.074, respectively. The curves for the four CM-Functions as well as the assumed one were shown in Figure 8, and the specific CMFs, bias as well as error percentages at some points are listed in Table 24. The CMFs derived from SPFs were overestimated (the safety benefits were underestimated) in all of the four sub-scenarios. When comparing the results between the two assumed CM-Functions for curve density, the bias and error percentage of Sub-Scenarios VI-3 and VI-4 (strong in curve density) were always much higher than those of VI-1 and VI-2 (weak in curve density), expect at a small range around 16. In short, the CMFs for curve density derived from SPFs were all biased when the relationship was nonlinear. The bias increased when the nonlinear level became stronger.

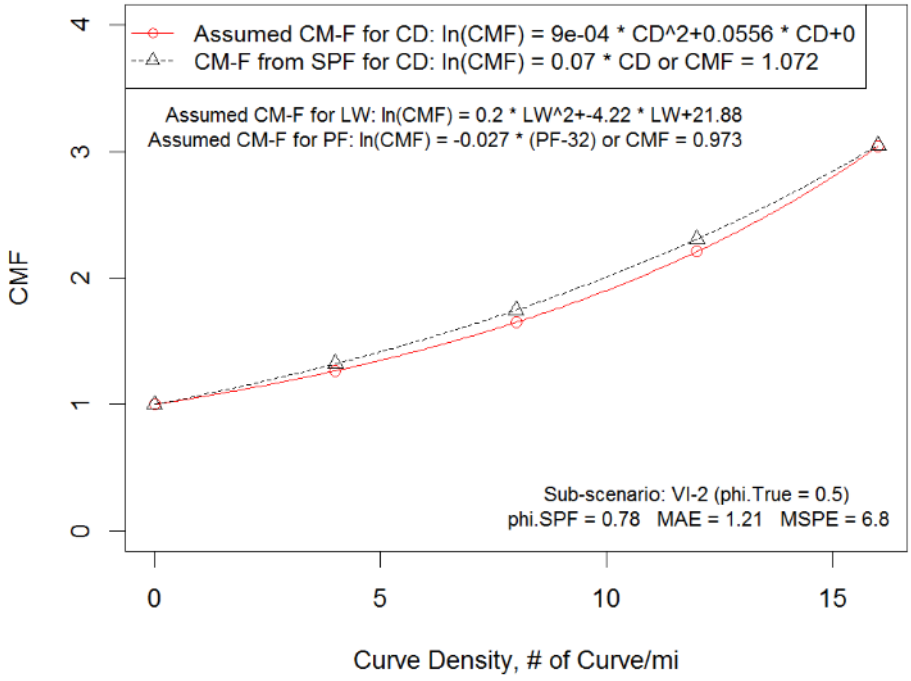
Another interesting finding was that the highest bias of CMF for curve density did not appear around the boundary areas, but near the middle. As can be seen in Figure

8, as curve density increased from the base point (i.e., zero), the bias first increased then decreased. The highest was around 11. This was probably due to the fact that the baseline for curve density was at the very left side. If the baseline was at some point in the middle (e.g., 8 or 10), the result might be similar to that of lane width. The bias should appear to be small around baseline and became large in boundary areas, intuitively. Nevertheless, the CMFs were still biased.

The CMFs for pavement friction produced from the four sub-scenarios with a 0.5 inverse dispersion parameter were 0.973, 0.972, 0.973 and 0.972, respectively. The curves for the assumed CM-Functions and those derived from SPFs in this scenario (with a 0.5 inverse dispersion parameter) are shown in Figure 9. Specific values of CMFs as well as bias and error percentage at some points are listed in Table 25. The overall results were nearly the same as those of Scenario V. Both bias and error percentage were small. The error percentages were all under 2 percent. So the CMFs for curve density were acceptable in all sub-scenarios.

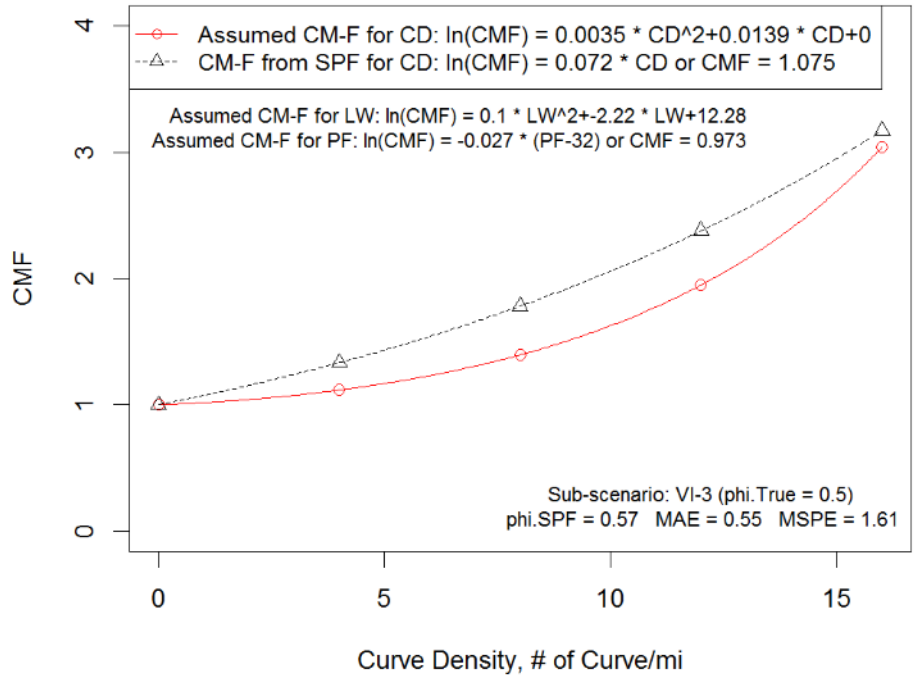


(a) Sub-Scenario VI-1

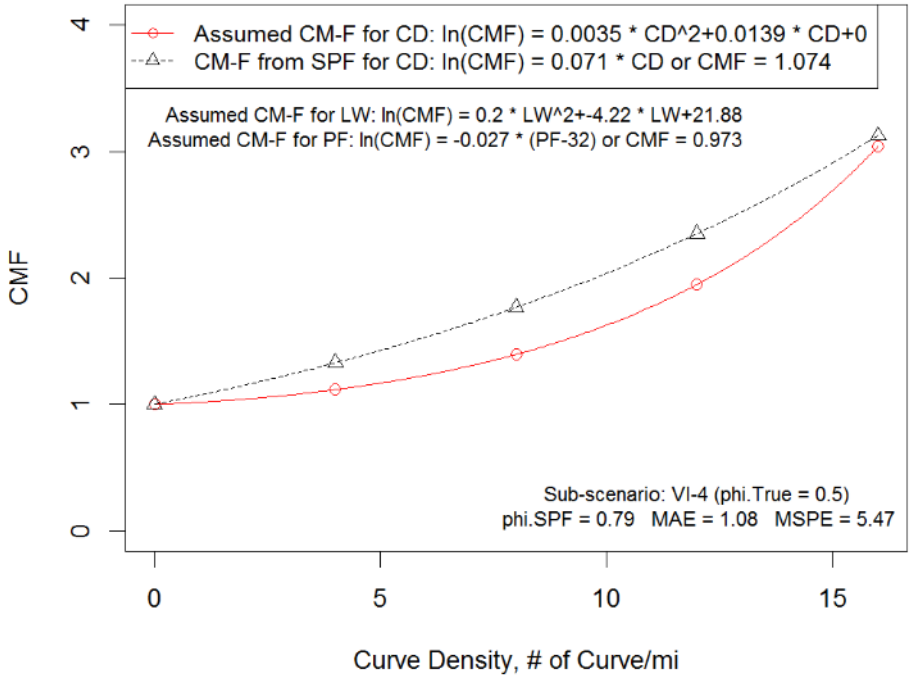


(b) Sub-Scenario VI-2

Figure 8 CM-Functions for Curve Density in Scenario VI (ϕ=0.5)



(c) Sub-Scenario VI-3



(d) Sub-Scenario VI-4

Figure 8 Continued

Table 24 Bias and Error of CMFs for Curve Density in Scenario VI

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
CD	4				8			
ϕ [d]=0.5								
VI-1	1.27	1.31	-0.04	3.5	1.65	1.72	-0.07	4.2
VI-2	1.27	1.30	-0.03	2.4	1.65	1.68	-0.03	1.9
VI-3	1.12	1.32	-0.20	18.0	1.40	1.74	-0.34	24.7
VI-4	1.12	1.31	-0.19	16.8	1.40	1.70	-0.31	22.0
ϕ [d]=1.0								
VI-1	1.27	1.31	-0.04	3.5	1.65	1.72	-0.07	4.1
VI-2	1.27	1.30	-0.03	2.4	1.65	1.68	-0.03	2.0
VI-3	1.12	1.32	-0.20	17.8	1.40	1.73	-0.34	24.1
VI-4	1.12	1.30	-0.19	16.6	1.40	1.70	-0.30	21.6
ϕ [d]=2.0								
VI-1	1.27	1.31	-0.04	3.4	1.65	1.71	-0.06	3.9
VI-2	1.27	1.29	-0.02	1.6	1.65	1.65	-0.01	0.3
VI-3	1.12	1.32	-0.20	17.7	1.40	1.73	-0.33	23.9
VI-4	1.12	1.31	-0.19	16.8	1.40	1.70	-0.31	22.1

Table 24 Continued

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
CD	12				16			
ϕ [d]=0.5								
VI-1	2.21	2.25	-0.04	2.0	3.04	2.95	0.09	2.9
VI-2	2.21	2.18	0.03	1.3	3.04	2.83	0.22	7.1
VI-3	1.95	2.30	-0.35	17.8	3.04	3.03	0.01	0.4
VI-4	1.95	2.22	-0.27	14.1	3.04	2.90	0.14	4.6
ϕ [d]=1.0								
VI-1	2.21	2.25	-0.04	1.9	3.04	2.95	0.09	3.0
VI-2	2.21	2.18	0.02	1.1	3.04	2.83	0.21	6.8
VI-3	1.95	2.28	-0.33	17.0	3.04	3.00	0.04	1.3
VI-4	1.95	2.21	-0.26	13.6	3.04	2.88	0.16	5.2
ϕ [d]=2.0								
VI-1	2.21	2.24	-0.03	1.6	3.04	2.94	0.10	3.4
VI-2	2.21	2.13	0.08	3.6	3.04	2.74	0.30	10.0
VI-3	1.95	2.28	-0.33	16.7	3.04	2.99	0.05	1.6
VI-4	1.95	2.23	-0.28	14.2	3.04	2.91	0.14	4.5

Note: a, b, c, d, * – the same notes as those in Table 14; CD = curve density (number of curves per mi).

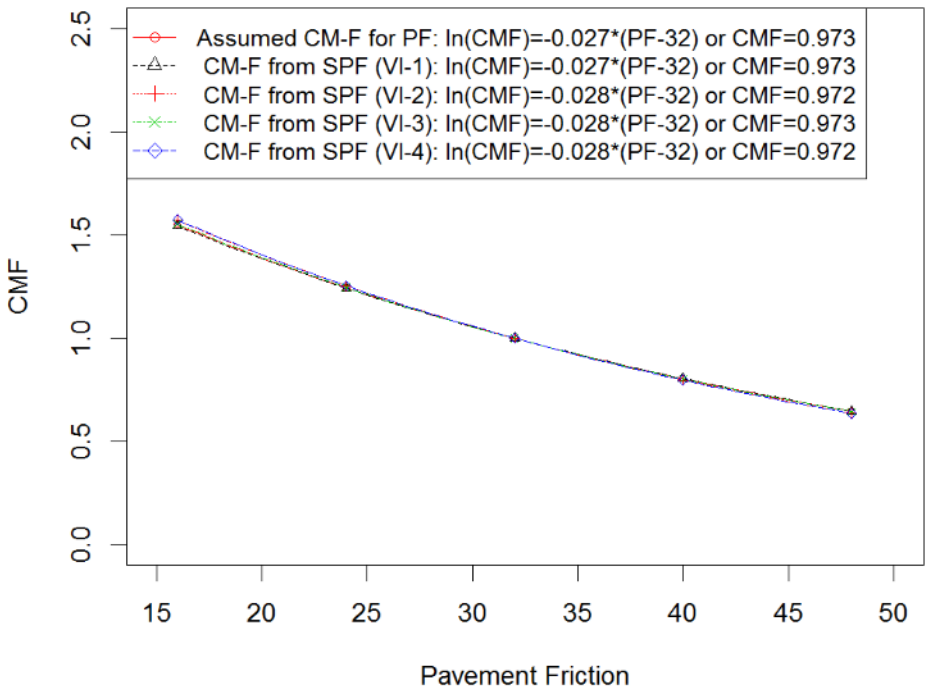


Figure 9 CM-Functions for Pavement Friction in Scenario VI (ϕ=0.5)

Table 25 Bias and Error of CMFs for Pavement Friction in Scenario VI

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
PF	16				24			
ϕ [d] = 0.5								
VI-1	1.55	1.56	-0.01	0.47	1.24	1.25	0.00	0.24
VI-2	1.55	1.55	0.00	0.06	1.24	1.24	0.00	0.03
VI-3	1.55	1.56	-0.01	0.52	1.24	1.25	0.00	0.26
VI-4	1.55	1.56	-0.01	0.49	1.24	1.25	0.00	0.25
ϕ [d] = 1.0								
VI-1	1.55	1.56	-0.01	0.63	1.24	1.25	0.00	0.31
VI-2	1.55	1.56	-0.01	0.38	1.24	1.25	0.00	0.19
VI-3	1.55	1.54	0.01	0.78	1.24	1.24	0.00	0.39
VI-4	1.55	1.56	-0.01	0.43	1.24	1.25	0.00	0.21
ϕ [d] = 2.0								
VI-1	1.55	1.56	-0.01	0.38	1.24	1.25	0.00	0.19
VI-2	1.55	1.53	0.02	1.12	1.24	1.24	0.01	0.56
VI-3	1.55	1.54	0.01	0.59	1.24	1.24	0.00	0.30
VI-4	1.55	1.57	-0.02	1.49	1.24	1.25	-0.01	0.74

Table 25 Continued

# *	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
PF	40				48			
ϕ [d]= 0.5								
VI-1	0.80	0.80	0.00	0.24	0.65	0.64	0.00	0.47
VI-2	0.80	0.80	0.00	0.03	0.65	0.65	0.00	0.06
VI-3	0.80	0.80	0.00	0.26	0.65	0.64	0.00	0.52
VI-4	0.80	0.80	0.00	0.25	0.65	0.64	0.00	0.49
ϕ [d]= 1.0								
VI-1	0.80	0.80	0.00	0.31	0.65	0.64	0.00	0.62
VI-2	0.80	0.80	0.00	0.19	0.65	0.64	0.00	0.38
VI-3	0.80	0.81	0.00	0.39	0.65	0.65	-0.01	0.79
VI-4	0.80	0.80	0.00	0.21	0.65	0.64	0.00	0.43
ϕ [d]= 2.0								
VI-1	0.80	0.80	0.00	0.19	0.65	0.64	0.00	0.37
VI-2	0.80	0.81	0.00	0.57	0.65	0.65	-0.01	1.14
VI-3	0.80	0.81	0.00	0.30	0.65	0.65	0.00	0.60
VI-4	0.80	0.80	0.01	0.74	0.65	0.64	0.01	1.46

Note: a, b, c, d, * – the same notes as those in Table 14; PF = pavement friction.

Recall that in Scenario V, CMFs for both curve density and pavement friction become less accurate as the nonlinear relationship between lane width and crash risk becomes strong, but this seems not to be always true to the CMF for pavement friction in this scenario. When the inverse dispersion parameter equaled to 0.5 and 1.0, CMFs for pavement friction in Sub-Scenario VI-3 (with weaker nonlinear relationship in lane width and stronger in curve density) had the highest bias and error percentage, and those in Sub-Scenario VI-2 (with stronger in lane width and weaker in curve density) had the lowest. But when the inverse dispersion parameter was 2.0, the former had the lowest while the latter had the highest. No possible reasons can be made at this moment.

The main findings of this scenario are close to those in Scenario V and can be summarized as follows: (1) the CM-Function for both lane width and curve density derived from SPFs were biased when using the common linear forms to model their nonlinear relationships; (2) with the increase of nonlinearity (i.e., nonlinear becomes stronger), the bias trended to become more significant; (3) the CMFs for other variables having linear relationship might be acceptable when mixed with those having nonlinear relationship; and (4) the misuse of linear link function for one or more variables led to biased estimate of other parameters.

The results in this section seem to contradict the results in Section 4.1 (i.e., linear relationships). In the previous section, the functional form used in regression models was of the same family with the one used to generate crash counts. The CMFs derived from SPFs were unbiased. In this section, they were not of the same family, and the CMFs were biased. It can be conclude from the two sections that functional forms play vital

roles in developing CMFs and/or CM-Functions in regression models, which is consistent with Hauer (2015). Unfortunately, the true safety effects of variables can be hardly known in practice, and this makes it difficult or impossible to identify the correct functional form in regression models. Safety analysts commonly adopt the linear form probably for its simplicity. This might be inadequate when some variables are having nonlinear effects on safety, as this section has illustrated.

4.4 Variable Correlation

Scenarios VII and VIII are discussed in this section.

4.4.1 Scenario VII: Variable Correlation, Linear Relationship

This scenario was essentially the same as Scenario I, except the two variables, AADT and lane width, were correlated in this scenario. The assumed CMF for lane width varied from 0.85 to 1.05 with an increment of 0.05. The theoretical function of the generated crash counts in this scenario is shown in Equation 4-1 (now as 4-19 below).

$$N_{true,i} = N_{spf,i} \times CMF_{LW,i} = 2.67 \times 10^{-4} \times L_i \times AADT_i \times exp[\beta_{LW} \times (LW_i - 12)] \quad (4\text{-}19a)$$

Or equivalently,

$$N_{true,i} = \beta_{off} \times L_i \times AADT_i \times exp(\beta_{LW} \times LW_i) \quad (4\text{-}19b)$$

The considered functional form is shown in Equation 4-2 (now as 4-20 below).

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\beta_2 \times LW_i) \quad (4\text{-}20)$$

The results are shown in Table 26. They were nearly the same as those of Scenario I. The estimation bias was relatively small under different simulation settings. The estimation bias was less than 0.02, and the error was within 2 percent.

Table 26 Results of Scenario VII

Theo. CMF [a]	CMF (SD) [b]	Bias	E [c]	AIC [d]	MAD [e]	MSPE [f]
ϕ = 0.5						
0.85	0.850 (0.033)	0.000	0.045	10181.9	0.038	0.010
0.90	0.903 (0.041)	0.003	0.322	10086.4	0.038	0.009
0.95	0.958 (0.039)	0.008	0.795	9994.2	0.036	0.008
1.00	1.002 (0.042)	0.002	0.205	9880.1	0.036	0.008
1.05	1.060 (0.042)	0.010	0.907	9785.6	0.035	0.007
ϕ = 1.0						
0.85	0.851 (0.047)	0.001	0.158	10315.9	0.050	0.015
0.90	0.905 (0.046)	0.005	0.508	10217.2	0.049	0.014
0.95	0.960 (0.054)	0.010	1.049	10101.4	0.046	0.014
1.00	1.006 (0.056)	0.006	0.578	9967.6	0.048	0.014
1.05	1.051 (0.052)	0.001	0.087	9877.4	0.044	0.013

Table 26 Continued

Theo. CMF [a]	CMF (SD) [b]	Bias	E [c]	AIC [d]	MAD [e]	MSPE [f]
ϕ = 2.0						
0.85	0.849 (0.051)	-0.001	0.111	10148.4	0.063	0.024
0.90	0.907 (0.068)	0.007	0.829	9998.2	0.061	0.022
0.95	0.943 (0.064)	-0.007	0.694	9892.5	0.058	0.020
1.00	1.017 (0.070)	0.017	1.729	9789.4	0.060	0.023
1.05	1.056 (0.064)	0.006	0.575	9727.2	0.053	0.018

Note: a – theoretical CMF; b – mean of CMFs from 100 experiments, SD is the Standard Deviation of the 100 CMFs; c – E is the error percentage, %; d, e, f – each is the mean value of the corresponding GOF measure of the 100 results.

Based on the result of this scenario, it seems the correlation between variables had trivial influence on CMFs derived from regression models.

4.4.2 Scenario VIII: Variable Correlation, Nonlinear Relationship

This scenario is the same as Scenario IV, except the dataset. The same three nonlinear CM-Functions were used for lane width, as shown in Equations 4-8, 4-9, and 4-10 (here as 4-21, 4-22, and 4-23 below), respectively.

$$ln(CMF) = 0.1 \times LW^2 - 2.22 \times LW + 12.28 \quad (4\text{-}21)$$

$$ln(CMF) = 0.2 \times LW^2 - 4.22 \times LW + 21.88 \quad (4\text{-}22)$$

$$ln(CMF) = \begin{cases} -0.11 \times (LW - 12)^2 + 0.30 & LW <= 12 \\ -0.08 \times (LW - 12)^2 + 0.30 & LW >= 12 \end{cases} \quad (4\text{-}23)$$

The nonlinear properties (closest line, area and AVD) for the three sub-scenarios are the same as those in Scenario IV (summarized in Table 12).

The theoretical function is shown in Equation 4-11 (here as 4-24 below).

$$N_{true,i} = N_{spf,i} \times CMF_{LW,i} = 2.67 \times 10^{-4} \times L_i \times AADT_i \times CMF_{LW,i} \quad (4\text{-}24)$$

The considered functional form is shown in Equation 4-12 (here as 4-25 below).

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp(\beta_2 \times LW_i) \quad (4\text{-}25)$$

Table 27 presents the CMFs derived from SPFs as well as other results (i.e., ϕ and GOF measurements) in this scenario. Figure 10 illustrates the curves of assumed CM-Functions for lane width and those derived from regression models in this scenario with a 0.5 inverse dispersion parameter. The specific CMFs for several lane widths of interest are listed in Table 28. First, the CMFs derived from the regression models were

all biased, especially around boundary areas. Second, Sub-Scenario VIII-2 (stronger nonlinearity) consistently had the highest MAD and MSPE. These two findings were consistent with that of Scenario IV. However, there were two significant differences between the two scenarios.

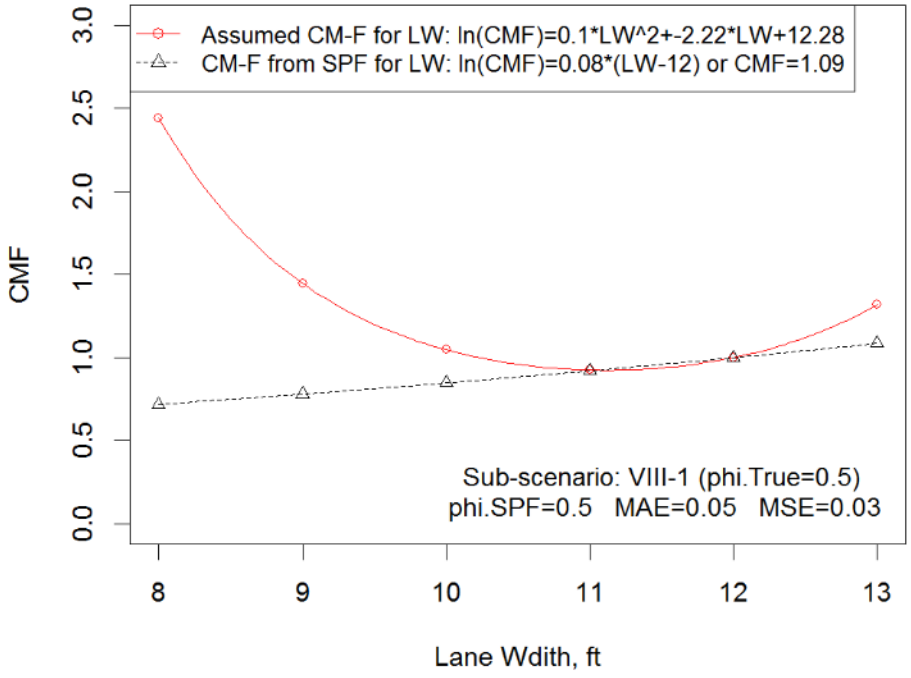
(1) The CMFs for lane width changed. The CMFs for lane width were 1.07, 1.49, and 1.12 in the three sub-scenarios, respectively, when the dispersion parameter was 0.5. They were obviously different with the corresponding ones in Scenario IV, which were 0.88, 0.98, and 1.33, respectively. Both the slope and intercept of the curve for CM-Function derived from SPFs changed when the variables became correlated, as shown in Figure 10. In this scenario, the two curves crossed at 11 and 12. Specifically, the bias and error percentage at 11-ft lane were really small, but they were relatively high at other points (i.e., 8-, 9-, and 10-ft lanes).

(2) The estimated dispersion parameter was close to the true value in this scenario, as shown in the third column of Table 27. Recall the findings in Scenario IV as well as Scenarios V and VII, the estimated dispersion parameter was biased when improper functional form was used. However, it was not in this scenario. That means the variables correlation also had significant influence on the estimates of other parameters in the regression models.

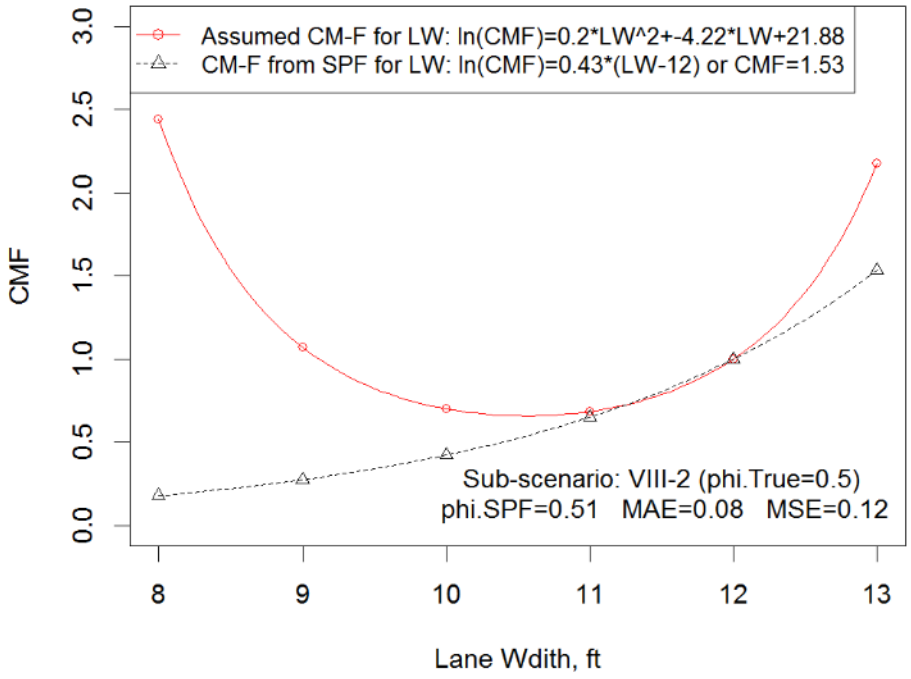
Table 27 Results of Scenario VIII

Sub-Scenario	CMF (SD) [a]	ϕ [b]	AIC [d]	MAD [e]	MSPE [f]
ϕ [c] = 0.5					
VIII-1	1.09 (0.05)	0.50	9845.45	0.05	0.03
VIII-2	1.53 (0.07)	0.51	9443.69	0.08	0.12
VIII-3	1.13 (0.05)	0.50	9624.95	0.05	0.02
ϕ [c] = 1.0					
VIII-1	1.07 (0.06)	0.99	9941.70	0.06	0.03
VIII-2	1.51 (0.09)	1.02	9520.91	0.09	0.14
VIII-3	1.12 (0.06)	1.00	9713.03	0.05	0.03
ϕ [c] = 2.0					
VIII-1	1.07 (0.08)	1.98	9758.17	0.07	0.04
VIII-2	1.48 (0.12)	2.03	9368.08	0.09	0.16
VIII-3	1.12 (0.07)	1.99	9504.21	0.07	0.03

Note: a – mean of CMFs from 100 experiments, SD is the Standard Deviation of the 100 CMFs; b - the inverse dispersion parameter derived from SPFs; c – the theoretical inverse dispersion parameter in each sub-scenario; d, e, f – each is the mean value of the corresponding GOF measure of the 100 results.

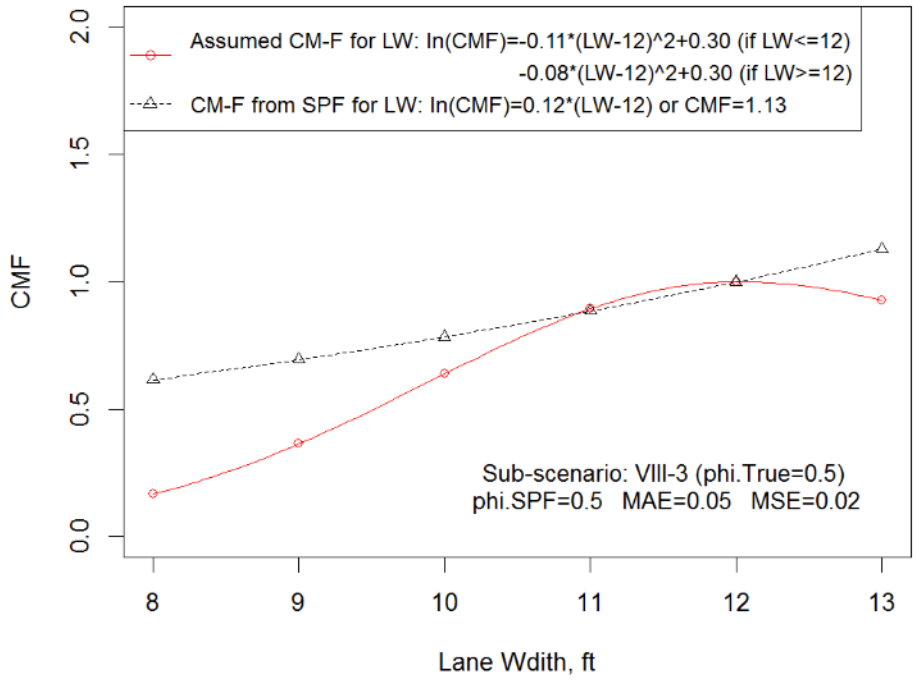


(a) Sub-Scenario VIII-1



(b) Sub-Scenario VIII-2

Figure 10 CM-Functions for Lane Width in Scenario VIII (ϕ=0.5)



(c) Sub-Scenario VIII-3

Figure 10 Continued

Table 28 Bias and Error of CMFs for Lane Width in Scenario VIII

Sub-Scenario	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
LW (ft)	8				9				10			
ϕ [d] = 0.5												
VIII-1	2.44	0.72	1.72	70.5	1.45	0.78	0.67	46.0	1.05	0.85	0.20	19.0
VIII-2	2.44	0.18	2.26	92.6	1.07	0.28	0.79	74.1	0.70	0.43	0.28	39.5
VIII-3	0.17	0.62	-0.45	271.0	0.36	0.69	-0.33	90.9	0.64	0.78	-0.15	22.9
ϕ [d] = 1.0												
VIII-1	2.44	0.75	1.69	69.3	1.45	0.81	0.64	44.3	1.05	0.87	0.18	17.3
VIII-2	2.44	0.19	2.25	92.1	1.07	0.29	0.78	72.8	0.70	0.44	0.26	37.3
VIII-3	0.17	0.63	-0.46	278.5	0.36	0.71	-0.34	93.7	0.64	0.79	-0.15	24.2
ϕ [d] = 2.0												
VIII-1	2.44	0.77	1.67	68.6	1.45	0.82	0.63	43.3	1.05	0.88	0.17	16.4
VIII-2	2.44	0.21	2.24	91.6	1.07	0.31	0.77	71.5	0.70	0.45	0.25	35.4
VIII-3	0.17	0.64	-0.48	288.0	0.36	0.72	-0.35	97.4	0.64	0.80	-0.16	25.7

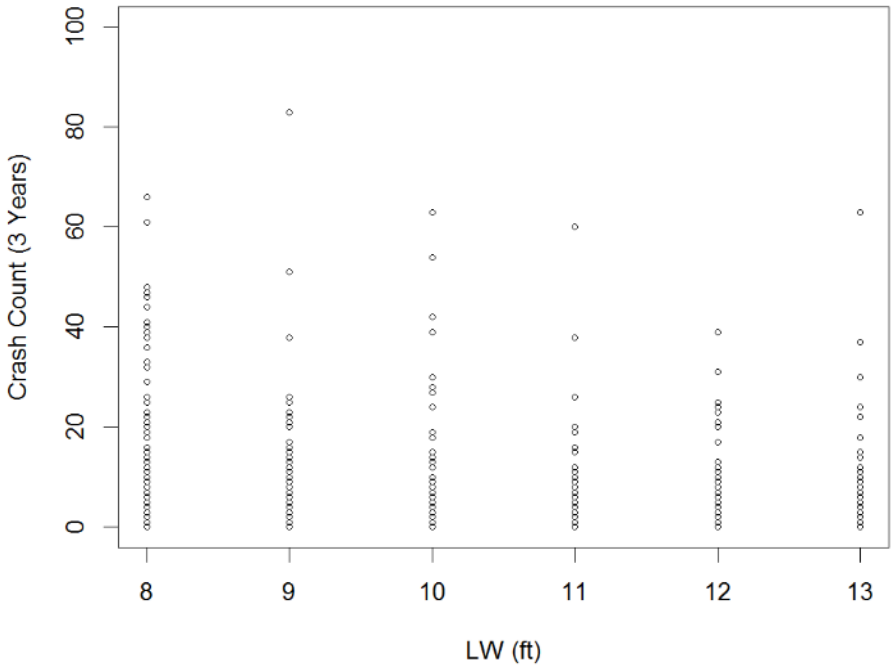
Table 28 Continued

Sub-Scenario	Th.[a]	SPF[b]	Bias	E[c]	Th.[a]	SPF[b]	Bias	E[c]
LW (ft)	11				13			
ϕ [d]= 0.5								
VIII-1	0.93	0.92	0.00	0.5	1.32	1.09	0.23	17.7
VIII-2	0.69	0.65	0.03	5.0	2.17	1.53	0.64	29.5
VIII-3	0.89	0.89	0.01	0.9	0.93	1.13	-0.20	21.8
ϕ [d]= 1.0								
VIII-1	0.93	0.93	0.00	0.5	1.32	1.07	0.24	18.5
VIII-2	0.69	0.66	0.02	3.3	2.17	1.51	0.67	30.7
VIII-3	0.89	0.89	0.00	0.4	0.93	1.12	-0.20	21.2
ϕ [d]= 2.0								
VIII-1	0.93	0.94	-0.01	1.1	1.32	1.07	0.25	19.0
VIII-2	0.69	0.67	0.01	1.8	2.17	1.48	0.69	31.7
VIII-3	0.89	0.90	0.00	0.2	0.93	1.12	-0.19	20.4

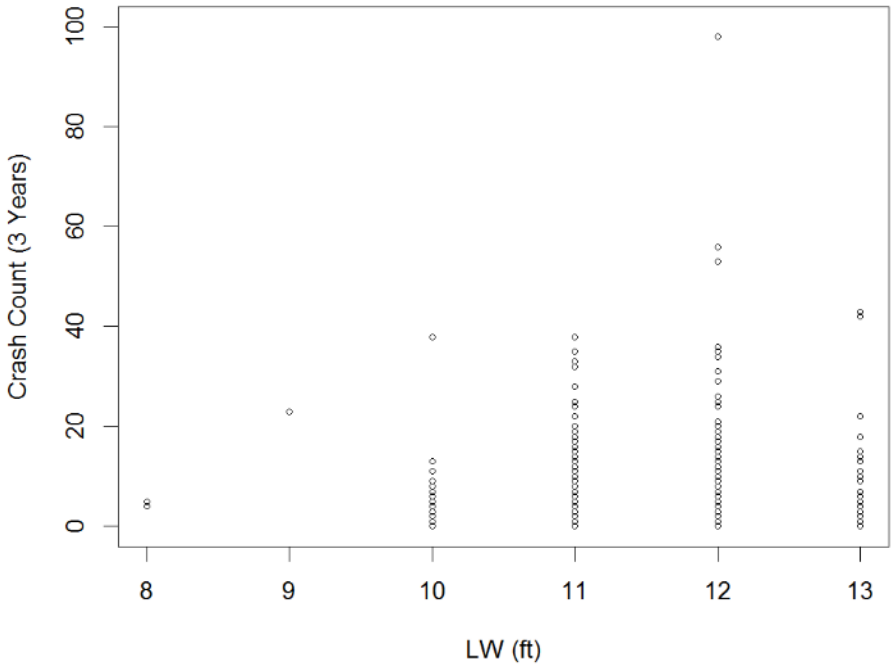
Note: a – theoretical CMF (assumed true specific CMFs for lane widths of 8, 9, 10, 11 and 12 ft); b – CMF derived from SPF (specific CMFs derived from regression models for corresponding lane widths); c – error percentage, %; d – the theoretical inverse dispersion parameter (ϕ) in each sub-scenario.

Another interesting finding is that the results differed between Scenarios VII and VIII. Variable correlation affected the CMFs derived from regression models when improper functions were used (Scenario VIII), but it did not when the functional form used in the model was of the same family with the one assumed (Scenario IV). Further analysis was conducted to analyze this phenomenon.

In order to examine the difference between the data points with variable correlation (i.e., Scenario VIII) and those without (i.e., Scenario IV), the scatter plots of crash counts against lane width in two sub-scenarios (IV-1 and VIII-1) are shown in Figure 11. It can be seen that, in Sub-Scenario IV-1, the points were nearly equally distributed among the five lane widths, whereas they were highly gathered together (11- and 12-ft lanes) in Sub-Scenario VIII-1. Calculation revealed that the five lane widths accounted for about 0.1 (8 ft), 0.1 (9 ft), 5.4 (10 ft), 33.2 (11 ft), 56.4 (12 ft), and 4.8 (13 ft) percent, respectively, in the latter sub-scenario. As can be seen, 11- and 12-ft lanes were prevalent among the five possible values. Together they accounted for nearly 90 percent. It is likely that the two groups of data points in Scenarios VII and VIII determined the regression results (i.e., the intercept and the slope).



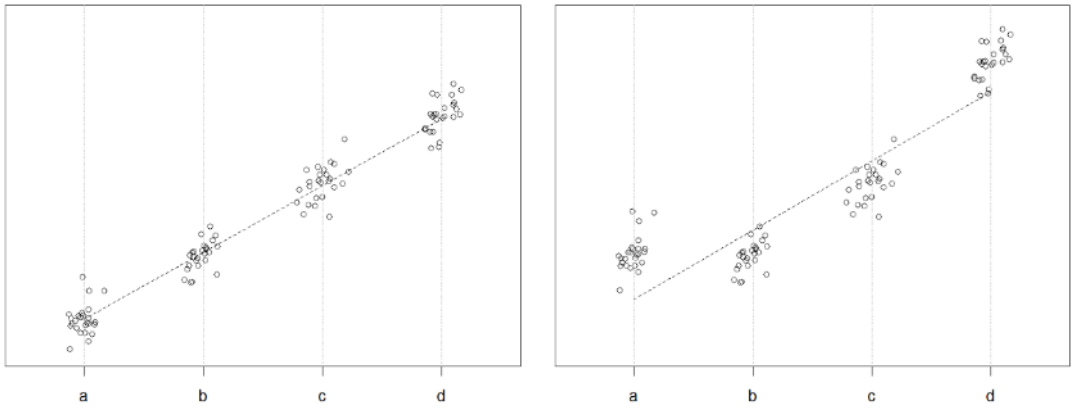
(a) Without Variable Correlation (Sub-Scenario IV-1)



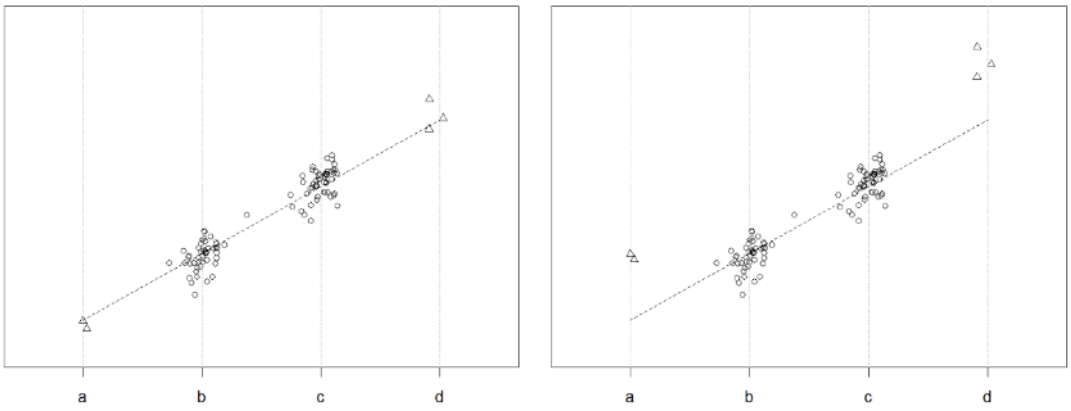
(b) With Variable Correlation (Sub-Scenario VIII-1)

Figure 11 Scatter Plots of Crash Counts and Lane Width

Figure 12 illustrated four types of dummy data points considering the distribution (uniform or non-uniform) and the relationship (linear or nonlinear). In (a) and (b) of Figure 12, the points were uniformly distributed, and the line were determined by the four groups of data points. In Figure 12 (a), the response variable and the explanatory variable had a linear relationship, so the line went through the four groups of data points roughly. In Figure 12 (b), they had a nonlinear relationship, then the line moved up to "better" fit the points, and of course biased. However, most of the data gathered around points b and c in Figure 12 (c) and (d). The intercept and slope of the line highly depended on these two groups of points. In Figure 12 (c), the response variable and the explanatory variable had a linear relationship, so the line again went through most of points. The relationship changed to nonlinear in Figure 12 (d), but the line did not move or rotate much. Because its intercept and slope were highly based on the two groups in the middle (i.e., points around b and c). The change of the data points on the two ends had trivial effect on the line since they were minority among all the data points. These four figures correspond to Scenarios I, IV, VII and VIII, respectively. This explains why variable correlations had no obvious effect in Scenario IV, but it influenced the CMFs significantly in Scenario VIII. Unfortunately, their effect on the estimation of other parameters (e.g., dispersion parameter) are not clear, and this needs further analysis in the future.



(a) Uniform and Linear (b) Uniform and Nonlinear



(c) Non-uniform and Linear (d) Non-uniform and Nonlinear

Figure 12 Example Illustrating the Effects of Variable Distribution on Regression

4.5 Combined Safety Effect

This sections discusses Scenario IX: combined safety effect. Two variables, lane width and shoulder width were considered in this scenario. The CMFs for them and other modeling results of each sub-scenario with a 0.5 inverse dispersion parameter are documented in Table 29. The results with other inverse dispersion parameters are presented in the Appendix. When compared with the previous scenarios, the bias and error percentage were relatively high in this scenario. The average error percentage was around 5.3% for CMFs of both lane width and shoulder width. The maximum was about 10%. When the adjustment factor was less than 1.0, the CMFs for both lane width and shoulder width were consistently underestimated. For example, the true CMFs for lane width and shoulder width were 0.8 and 0.85, respectively, in Sub-Scenario IX-1 (adjustment factor equaled to 0.80). Those derived from regression models were 0.73 and 0.77, respectively. Safety analysts may misleadingly overestimate the safety benefits of widening the lane and that of widening the shoulder. The results were contrary when the adjustment factor was more than 1.0. CMFs were overestimated and benefits of widening lane or shoulder individually were both underestimated. So, neither the CMFs for lane width nor those for shoulder width can reflect their true individual safety effectiveness in this scenario.

Table 29 Results of CMFs in Scenario IX (ϕ = 0.5)

# *	AF [a]	LW [b]				SW [b]				ϕ [b]	AIC[d]	MAD[e]	MSPE[f]
		Th.	SPF (SD)	Bias	E	Th.	SPF (SD)	Bias	E				
IX-1	0.80	0.8	0.73 (0.048)	-0.07	9.26	0.85	0.77 (0.046)	-0.08	8.92	0.49	8750.40	0.052	0.014
IX-2	0.90	0.8	0.77 (0.047)	-0.03	4.06	0.85	0.81 (0.048)	-0.04	4.19	0.50	8908.33	0.043	0.012
IX-3	0.95	0.8	0.79 (0.049)	-0.01	1.46	0.85	0.83 (0.046)	-0.02	1.77	0.49	8917.64	0.040	0.010
IX-4	1.05	0.8	0.82 (0.05)	0.02	2.57	0.85	0.88 (0.05)	0.03	3.17	0.50	9056.37	0.039	0.010
IX-5	1.10	0.8	0.83 (0.054)	0.03	3.77	0.85	0.89 (0.054)	0.04	4.44	0.49	9106.98	0.044	0.012
IX-6	1.20	0.8	0.87 (0.047)	0.07	8.72	0.85	0.93 (0.052)	0.08	9.20	0.50	9200.09	0.051	0.014
IX-7	0.80	0.9	0.82 (0.043)	-0.08	9.08	0.85	0.77 (0.049)	-0.08	9.68	0.50	9046.82	0.055	0.016
IX-8	0.90	0.9	0.86 (0.046)	-0.04	4.50	0.85	0.8 (0.044)	-0.05	5.55	0.49	9162.17	0.043	0.011
IX-9	0.95	0.9	0.88 (0.049)	-0.02	2.08	0.85	0.83 (0.048)	-0.02	2.64	0.50	9223.60	0.042	0.011

Table 29 Continued

# *	AF [a]	LW [b]				SW [b]				ϕ [b]	AIC[d]	MAD[e]	MSPE[f]
		Th.	SPF (SD)	Bias	E	Th.	SPF (SD)	Bias	E				
IX-10	1.05	0.9	0.92 (0.045)	0.02	2.57	0.85	0.86 (0.045)	0.01	1.76	0.49	9307.60	0.040	0.009
IX-11	1.10	0.9	0.94 (0.055)	0.04	4.11	0.85	0.89 (0.049)	0.04	4.80	0.50	9366.01	0.043	0.010
IX-12	1.20	0.9	0.98 (0.06)	0.08	9.04	0.85	0.93 (0.05)	0.08	9.95	0.50	9451.37	0.054	0.015
IX-13	0.80	0.8	0.73 (0.035)	-0.07	9.00	0.9	0.82 (0.047)	-0.08	8.48	0.49	8927.76	0.054	0.016
IX-14	0.90	0.8	0.76 (0.046)	-0.04	5.51	0.9	0.87 (0.047)	-0.03	3.66	0.50	9003.28	0.042	0.011
IX-15	0.95	0.8	0.78 (0.04)	-0.02	2.92	0.9	0.88 (0.047)	-0.02	2.19	0.49	9072.50	0.038	0.009
IX-16	1.05	0.8	0.81 (0.045)	0.01	1.65	0.9	0.92 (0.055)	0.02	2.31	0.49	9184.27	0.041	0.009
IX-17	1.10	0.8	0.84 (0.046)	0.04	4.45	0.9	0.94 (0.051)	0.04	4.11	0.49	9232.53	0.041	0.009
IX-18	1.20	0.8	0.87 (0.057)	0.07	9.33	0.9	0.98 (0.054)	0.08	8.58	0.49	9330.62	0.053	0.014

Table 29 Continued

#	AF [a]	LW [b]				SW [b]				ϕ [c]	AIC[d]	MAD[e]	MSPE[f]
		Th.	SPF (SD)	Bias	E	Th.	SPF (SD)	Bias	E				
IX-19	0.80	0.9	0.81 (0.05)	-0.09	9.75	0.9	0.81 (0.048)	-0.09	9.64	0.51	9155.48	0.058	0.018
IX-20	0.90	0.9	0.86 (0.044)	-0.04	4.74	0.9	0.87 (0.047)	-0.03	3.64	0.50	9287.17	0.044	0.011
IX-21	0.95	0.9	0.88 (0.047)	-0.02	2.13	0.9	0.88 (0.053)	-0.02	2.18	0.50	9358.16	0.042	0.010
IX-22	1.05	0.9	0.92 (0.058)	0.02	2.70	0.9	0.93 (0.055)	0.03	3.43	0.51	9447.52	0.044	0.012
IX-23	1.10	0.9	0.95 (0.049)	0.05	5.29	0.9	0.94 (0.053)	0.04	4.60	0.49	9490.10	0.044	0.011
IX-24	1.20	0.9	0.99 (0.061)	0.09	9.80	0.9	0.99 (0.054)	0.09	10.12	0.50	9627.21	0.056	0.016

Note: # = Sub-scenario number; a – AF is the assumed adjustment factor; b – LW is for lane width, SW is for shoulder width, Th. means the true CMF value, SPF is the mean of CMFs from 100 experiments, SD is the standard deviation of the 100 CMFs, E is error percentage (%); c - the mean of inverse dispersion parameter estimated from 100 experiments; d, e, f – each is the mean of the corresponding GOF of the 100 results.

Further, the relationship between the accuracy of CMFs and the presumed adjustment factors were investigated. The relationship between error percentage and adjustment factor are illustrated in Figures 13 and 14. Figure 13 shows the error percentage of CMFs for lane width and Figure 14 shows that for shoulder width. The two figures clearly indicate that the error percentage was highly related to the adjustment factor. The error percentage was consistently the highest when the adjustment factor was 0.80 or 1.20. And the lowest when it was 0.95 or 1.05. The error percentage became small as the adjustment factor became closer to 1.0. A special case can be seen when the adjustment factor equaled to 1.0, the scenario configuration fell into that in Scenario II (with two variables). The error percentage should be much lower (technically zero) based on the findings in that scenario. So the adjustment factor considerably influenced the CMFs for both lane width and shoulder width. When it was close to 1.0, this influence might be minor. But when it became far from 1.0 (i.e., less than or more than 1.0), the accuracy of CMFs can be significantly affected. The further away it is from 1.0, the lower the quality of the CMFs is. In other words, the CMFs were biased when the multiple treatments were actually not affecting crash risk independently. The rate at which the value became biased was actually very high when the adjustment factor went away from 1.0.

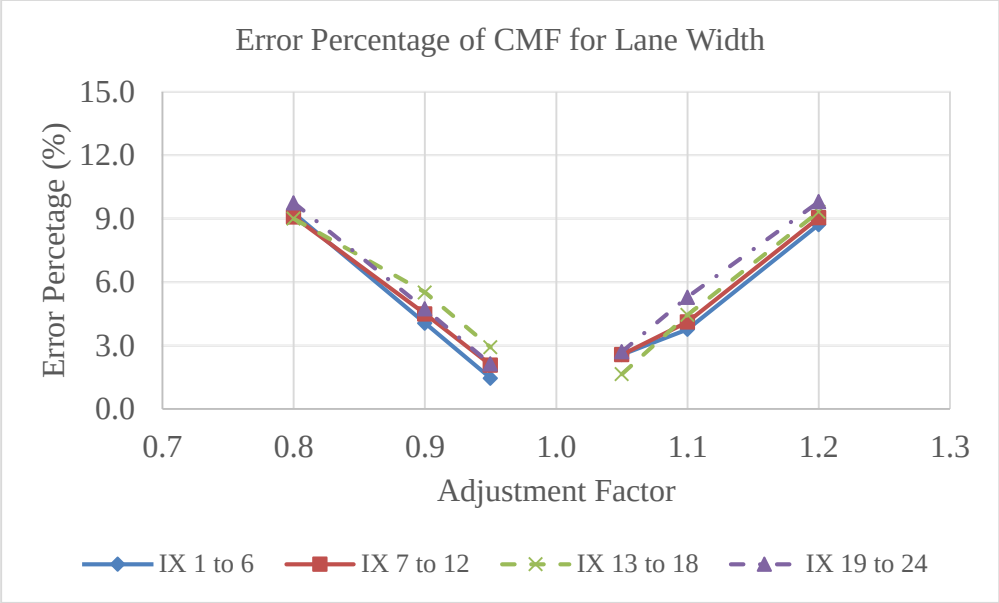


Figure 13 Error Percentage of CMFs for Lane Width in Scenario IX (ϕ=0.5)

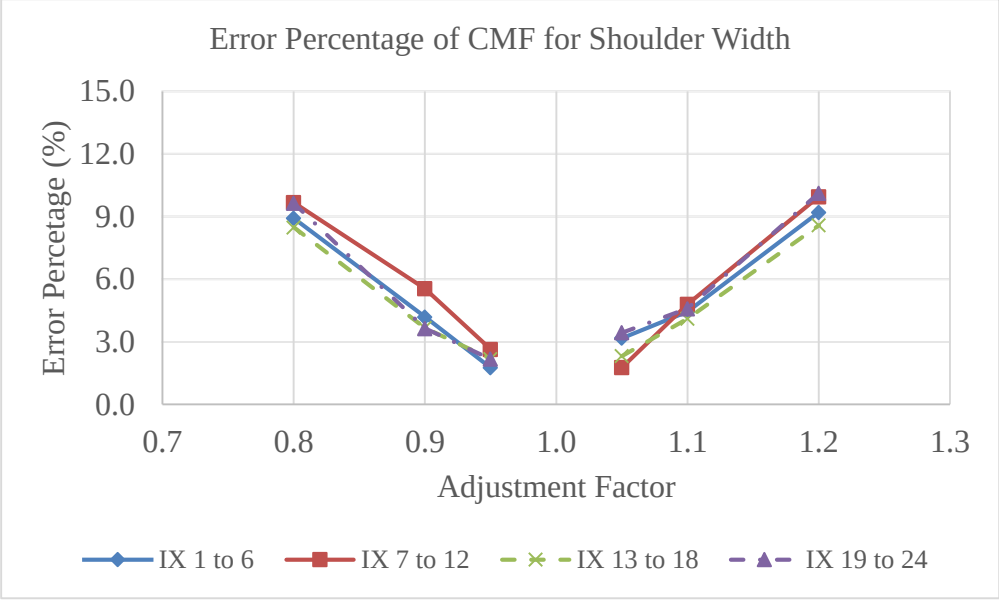


Figure 14 Error Percentage of CMFs for Shoulder Width in Scenario IX (ϕ=0.5)

The row of "ϕ" in Table 29 lists the estimated inverse dispersion parameters from the regression models for each scenario. All of them were very close to the corresponding true values regardless of the assumed CMFs for variables. No significant influence of the adjustment factor on the estimate of inverse dispersion parameters was found in this scenario. Similar results were observed for other inverse dispersion parameters.

Another interesting finding from this scenario was the GOF measurements. Both MAE and MSPE were relatively small in each sub-scenario, and they were very close to those in Scenario I. This indicated that the predicated crash number was quite close to the true crash mean. However, this did not guarantee the quality of CMFs derived from regression models, as has been described above. That is to say, although the fitting result seems to be good in terms of GOF measurements, there can still be some substance issues with the models. A possible reason is that some parameters may have been overestimated (or underestimated) while others may have been underestimated (or overestimated) in the regression models. Take Sub-Scenario IX-1 as an example. The specific theoretical function for generating crash counts is shown in Equation 4-26.

$$N_{true,i} = 2.67\times10^{-4} \times L_i \times AADT_i \times 0.8^{LW_i-12} \times 0.85^{SW_i-6} \times AF^{I_{\{LW_i\neq12\}}(LW_i)I_{\{SW_i\neq6\}}(SW_i)}$$

(4-26a)

Or equivalently,

$$N_{true,i} = 0.103\times AF^{I_{\{LW_i\neq12\}}(LW_i)I_{\{SW_i\neq6\}}(SW_i)} \times L_i \times AADT_i \times exp(-0.22LW_i - 0.16SW_i)$$

(4-26b)

Where,

$I_{\{LW_i \neq 12\}}(LW_i)$ = indicator function for lane width of segment *i*. It equals to 0 if the lane width is 12 ft, otherwise 1.0, and,

$I_{\{SW_i \neq 6\}}(SW_i)$ = indicator function for shoulder width of segment *i*. It equals to 0 if the shoulder width is 6 ft, otherwise 1.0.

The modeling output of one experiment in this sub-scenario is shown in Table 30. It can be seen that the coefficients for lane width and shoulder width were both obviously underestimated. And that for AADT was slightly underestimated. But the intercept coefficient was overestimated. Note that the specific theoretical value for the intercept is not directly given in Table 30 due to the fact that it depends on the two indicator functions. In other words, the theoretical intercept varied when the segment group changed. For segment Groups 1, 2 and 3, it was -2.27 (logarithm of 0.103). But it was -2.49 (logarithm of the product of 0.103 and AF, 0.8) for segment Group 4. The coefficient estimated from regression models was much higher than either of them. In this experiment, the coefficients for lane width and shoulder width were both underestimated. It seems the underestimation was compensated through overestimating the intercept coefficient. Perhaps this explains the overall smaller MAE and MSPE values.

Table 30 Modeling Output of the an Experiment in Sub-Scenario IX-1

Model Variable	Theo. Value [a]	Coef. Value [b]	SE [c]	p-Value
Intercept [$ln(\beta_0)$]	-2.27-0.223$I_{LW}I_{SW}$ [d]	-1.810	0.713	0.0111
Ln(AADT) (β_1)	1.00	0.981	0.040	< 2e-16
Lane Width (β_2), ft	-0.223	-0.352	0.045	3.26E-15
Shoulder Width (β_3), ft	-0.162	-0.321	0.045	6.18E-13
AIC			8921.8	
MAD			0.050	
MSPE			0.014	

Note: a – theoretical value; b – estimated coefficient value; c – SE is standard error;
d - the theoretical value for intercept was calculated by taking the nature logarithm of the first two terms of Equation 4-26b, $0.103\times AF^{I_{\{LW_i\neq 12\}}(LW_i)I_{\{SW_i\neq 6\}}(SW_i)}$. I_{LW} and I_{SW} are the two indicator functions of lane width and shoulder width, respectively.

4.6 Summary

This chapter has described the detailed evaluation results of CMFs derived from regression models using simulated crash data with nine scenarios. The simulation has shown several key findings.

(1) Scenarios I and II considered linear relationships between variables and crash risk and assumed all the requirements of a cross-sectional study were satisfied. The result indicated the CMFs produced using the common regression models should be unbiased under such conditions.

(2) Scenario II focused on the omitted-variable problem. Simulation analyses confirmed that regression models suffer from this problem, and the CMFs derived from the models were biased when some factors having significant effect on safety were omitted. When this effect was minor, the quality of CMFs might be acceptable.

(3) Scenarios III, IV and VI examined the bias of CMFs when some variables had nonlinear relationship with crash risk. It was found that link functions were crucial in developing reliable CMFs. The commonly used GLMs were likely to produce biased CMFs when the relationship between variables and crash risk were not linear (in logarithm format). In addition, this also led to biased estimates for other parameters.

(4) Scenarios VII and VIII were repetitions of Scenarios I and IV, respectively, with emphasis on variable correlation. It was found that the correlation between variables had no obvious influence on the CMFs under conditions of linear relationships. However, it affects the CMFs significantly when improper functional form was used in the regression models.

(5) Scenario IX considered the independence assumption within the common regression models. Once this assumption was not met, the individual CMFs for multiple variables or treatments included in the regression models were biased, especially when the dependence was strong.

To verify the simulation findings, observed data was analyzed and the results are discussed in the next chapter.

5. VALIDATION USING OBSERVED DATA*

In previous chapters, simulation had been used to examine the quality of CMFs developed from regression models. The simulation analyses mainly raised three problems with the use of regression models for estimating CMFs. Specifically, they are the omitted-variable problem, functional form, and dependence of variables. In order to better illustrate the findings from the simulation analyses, the CMFs were derived from an observed dataset with regression methods. Due to the fact that little was known about the combined safety effects of multiple variables, it was not easy to validate the variable dependent problem. Thus, this chapter mainly considered the functional form and omitted-variable problems.

Section 5.1 briefly introduces the dataset used in this study. Section 5.2 presents the modeling results and CM-Functions derived from regression models with various functional forms. Section 5.3 documents the volume-only model and "full-variable" model. Section 5.4 provides a summary of the work accomplished with the observed dataset.

5.1 Data Description

The real observed dataset was requested from the Highway Safety Information System (HSIS) managed by the FHWA (FHWA 2011). Segments of two-lane rural

* The real crash date used in this chapter was provided by the Highway Safety Information System (HSIS). The author greatly appreciates HSIS for providing the data.

highways in Washington State was identified, and three-year (from 2006 to 2008) crash records on these segments were collected. The segment length, traffic volume (i.e., AADT), lane width and shoulder width (right + left) of each segment were obtained. Note that the AADT was the average in the three years. The segments with AADT less than 500 were excluded, because some agencies pointed out the AADT of those segments were usually unreliable (Srinivasan and Carter 2011). In addition, some segments with obvious mistakes (e.g., 0-ft or extreme wide lanes) were removed. Finally, 8,132 segments were identified. The segments are summarized in Table 31. The scatter plots of crash rate (number of crashes per mile) against the three variables (i.e., AADT, lane width, and shoulder width) are illustrated in Figure 15.

Table 31 Summary Statistics of Observed Data

Variable	Sample Size	Min.	Max	Mean (SD *)
Length (mi)	8132	0.01	7.51	0.3 (0.5)
AADT	8132	511	26856	4263.4 (3948.4)
Lane Width (ft)	8132	9	14	11.6 (0.7)
Shoulder Width (ft)	8132	0	44	9.6 (5.4)
Crash Count (3 years)	8132	0	50	1.4 (2.6)

Note: * SD = standard deviation.

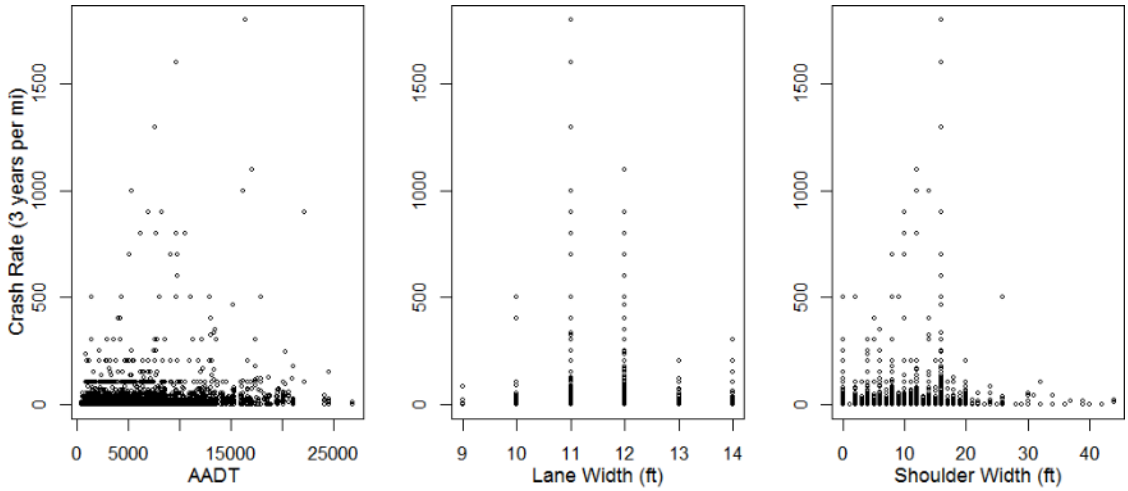


Figure 15 Scatter Plots of Crash Rate against Variables

5.2 Estimated CMFs using Different Functional Forms

Based on the simulation results, functional forms played important roles in developing CMFs. The use of improper functional forms could lead to significant bias to CMFs developed from regression models. To validate the importance of functional forms, GLM and GNM were used to analyze the real observed data. Six types of link functions were utilized in the regression models. They were linear, inverse, exponential, log, power, and quadratic functions, respectively. The model with linear link function was equivalent to the commonly used GLM. The other five link functions were adopted from a recent study (Park and Abdel-Aty 2015b). The modeling procedure of GNMs was generally the same as the previous studies conducted by Lao et al. (2014) and Lee et al. (2015), but necessary improvements were made. In the previous studies, the nonlinear link functions were determined by observing the curve pattern between crash rate (in

logarithm form) and the variables of interests. This could lead to bias, because the link functions were subject to observers. Instead, this study developed the nonlinear link functions, given the forms, by fitting the relationship between crash rate and variable (i.e., lane width). It is worth to mention that this section only considered lane width in the regression models to simplify the analysis, the result of which might be affected by the omitted-variable bias. However this would not influence the objective of the analysis.

Prior to regression analyses, the relationship between crash rate and lane width was explored. The curve is shown in Figure 16. The vertical axis in Figure 16 is mean crash rate (number of crashes in three years per mile per one thousand AADT), and the horizontal axis represents lane width. It can be seen that the overall crash rate decreased as the lane width increased. However, the decreasing rate was obviously higher when the lane was narrower. Widening segments with narrower lanes seemed to be more effective than widening wider ones if other factors (e.g., AADT) were the same or similar. This indicated a linear function might not be adequate to reflect the relationship between the two.

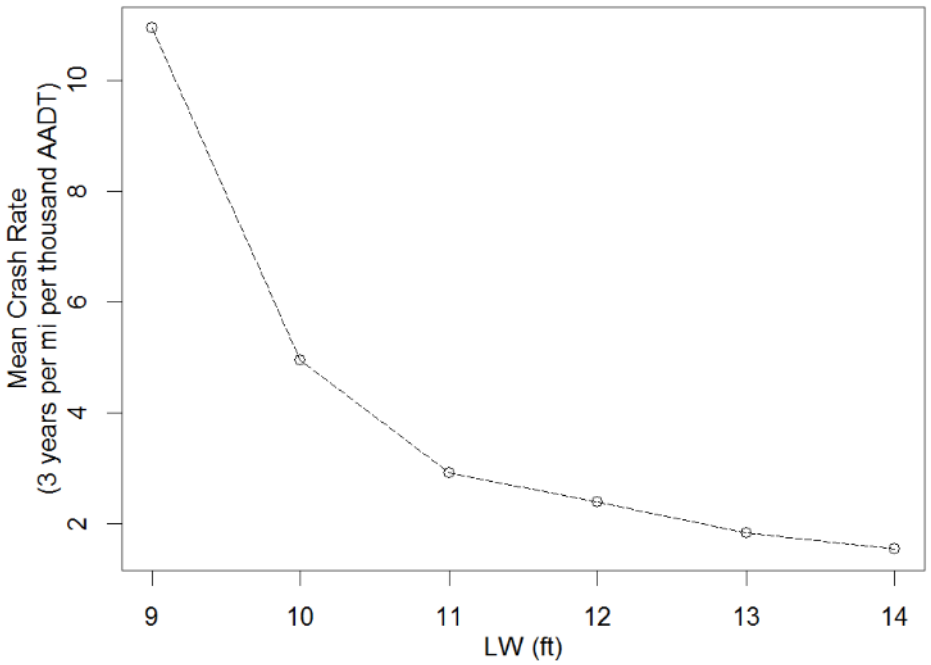


Figure 16 Mean Crash Rate and Lane Width

5.2.1 Linear

The linear functional form assumed the relationship between crash rate (in logarithm) and variable of interest (i.e., lane width) followed Equation 5-1.

$$U(LW) = \ln(CR) = A + B \times LW \quad (5\text{-}1)$$

Where,

$U(LW)$ = link function for lane width;

CR = crash rate;

LW = lane width (ft); and,

A and B are coefficients to be estimated.

Actually, this was consistent with the commonly used GLMs. The target model form was the same as Equation 4-2 (now as 5-2 below).

$$\beta \qquad _2 \qquad (5\text{-}2)$$

The modeling results (coefficient estimates and GOF measures) are presented in Table 32. All the three parameters were statistically significant at a 99 percent level. The CMF for lane width derived from this model was 0.85 ($e^{-0.16}$), meaning the expected crash number would decrease by 15 percent whenever the lane was widen by 1 foot. The CM-Function is illustrated in Figure 17.

Table 32 Modeling Result of Observed Date with Linear Functional Form

Model Variable	Coef. Value [a]	SE [b]	p-Value
Intercept [$ln(\beta_0)$]	-4.72	0.31	5.84E-53
Ln(AADT) (β_1)	1.02	0.02	0.00E+00
Lane Width (β_2), ft	-0.16	0.03	7.58E-10
ϕ	1.3344	0.0542	-
AIC		22083.9	
MAD		1.211	
MSPE		5.418	

Note: a – estimated coefficient value; b – SE = standard error.

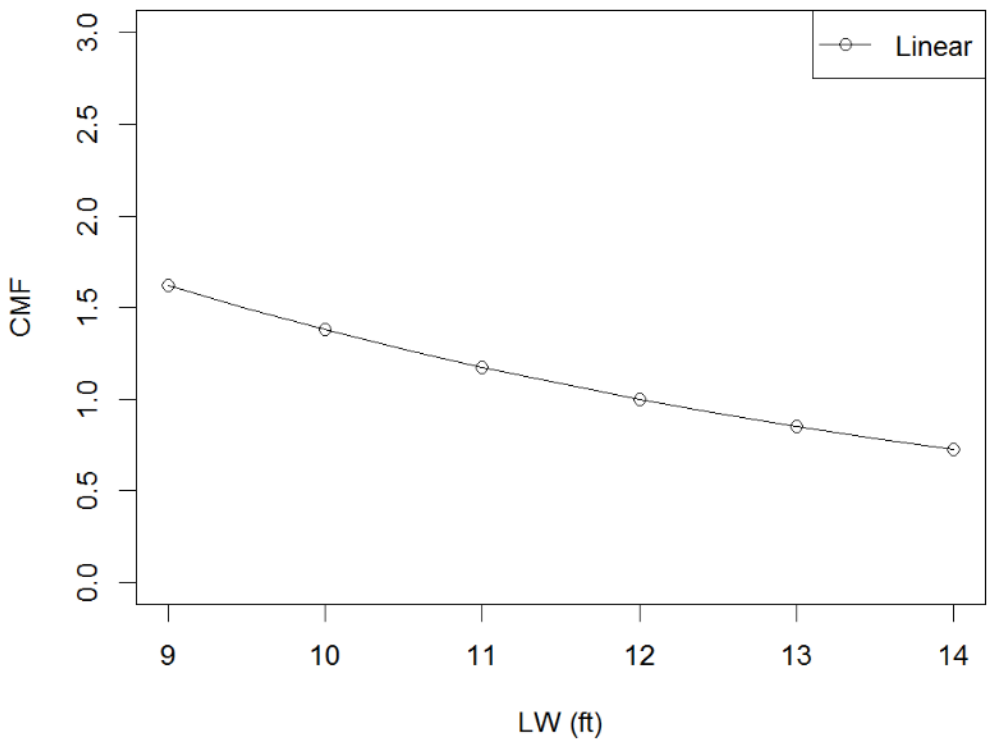


Figure 17 CM-Function for Lane Width Derived using Observed Data (Linear)

5.2.2 Inverse

With inverse functional form, the assumed relationship between crash rate (in logarithm) and lane width is shown in Equation 5-3.

$$U(LW) = \ln(CR) = A + B / LW \quad (5\text{-}3)$$

The fitting results of inverse link function is shown in Table 33 and the specific function is shown in Equation 5-4. Both A and B were statistically significant at a 99 percent level.

$$U(LW) = -0.0063 + 0.1037 / LW \quad (5\text{-}4)$$

Table 33 Fitting Result of Inverse Functional Form for Lane Width

Model Variable	Coef. Value [a]	SE [b]	p-Value
A	-0.0063	0.0020	1.92e-03
B	0.1037	0.0234	9.62e-06

Note: a – estimated coefficient value; b – SE = standard error.

The target model form is shown in Equation 5-5.

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp[\beta_2 \times U(LW_i)] \quad (5\text{-}5)$$

In Equation 5-5, the link function for lane width, $U(LW)$, was substituted by Equation 5-4. The final results are shown in Table 34.

Table 34 Modeling Result of Observed Date with Inverse Functional Form

Model Variable	Coef. Value [a]	SE [b]	p-Value
Intercept [$ln(\beta_0)$]	-7.17	0.20	< 2e-16
Ln(AADT) (β_1)	1.02	0.02	< 2e-16
U(LW) (β_2)	211.94	33.11	1.54e-10
ϕ	1.3353	0.0543	-
AIC		22081.5	
MAD		1.210	
MSPE		5.411	

Note: a – estimated coefficient value; b – SE = standard error.

The CM-Function for lane width was then estimated as Equation 5-6.

$$ln(CMF) = \beta_2 \times U(LW) - \beta_2 \times U(12) = 211.94 \times 0.1037(\frac{1}{LW} - \frac{1}{12}) \quad (5\text{-}6a)$$

Or equivalently,

$$CMF = exp[21.97 \times (\frac{1}{LW} - \frac{1}{12})] \quad (5\text{-}6a)$$

The value of 12 in Equation 5-6 reflects the base condition for lane width. The CM-Function for lane width with inverse functional form is illustrated in Figure 18.

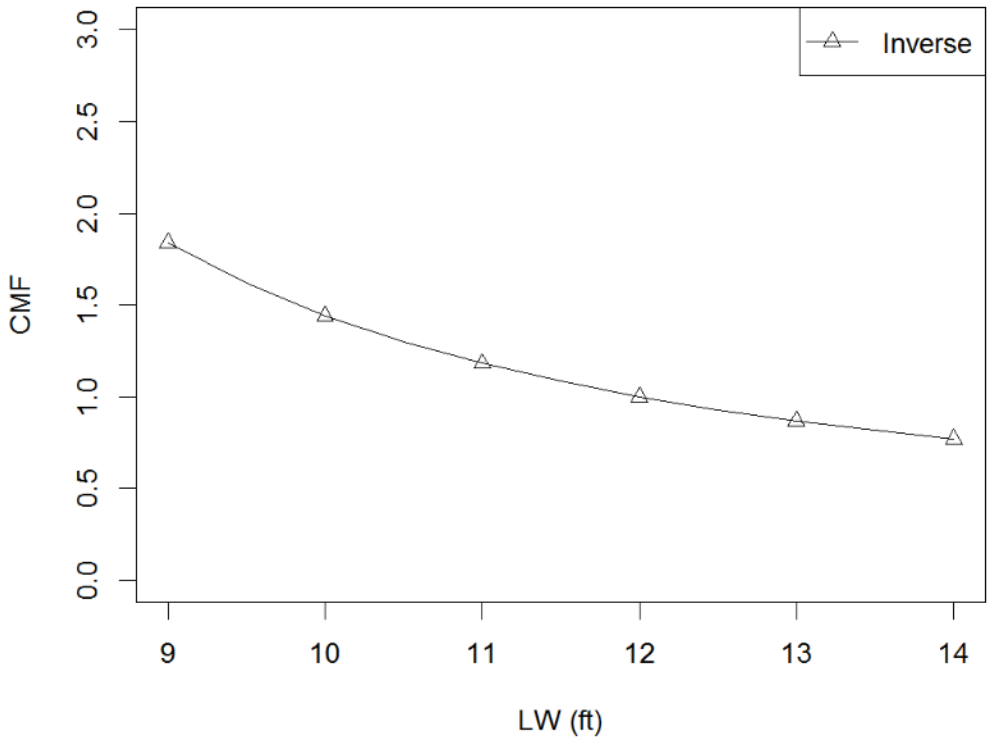


Figure 18 CM-Function for Lane Width Derived using Observed Data (Inverse)

5.2.3 Exponential

The exponential form is shown in Equation 5-7.

$$U(LW) = \ln(CR) = A + B \times exp(LW) \quad (5\text{-}7)$$

Similar procedures were used as that for the inverse form. The results are shown in Tables 35 and 36.

Table 35 Fitting Result of Exponential Functional Form for Lane Width

Model Variable	Coef. Value [a]	SE [b]	p-Value
A	2.972e-03	1.644e-04	< 2e-16
B	-2.111e-09	8.012e-10	8.43E-03

Note: a – estimated coefficient value; b – SE = standard error.

Table 36 Modeling Result of Observed Date with Exponential Functional Form

Model Variable	Coef. Value [a]	SE [b]	p-Value
Intercept [$ln(\beta_0)$]	-7.12	0.26	< 2e-16
Ln(AADT) (β_1)	1.01	0.02	< 2e-16
U(LW) (β_2)	239.14	69.83	6.16E-04
ϕ	1.3271	0.0538	-
AIC		22107.4	
MAD		1.214	
MSPE		5.415	

Note: a – estimated coefficient value; b – SE = standard error.

The CM-Function for lane width is shown in Equation 5-8. And the curve is illustrated in Figure 19.

$$ln(CMF) = \beta_2 \times U(LW) - \beta_2 \times U(12) = -5.05 \times 10^{-7} \times [exp(LW) - exp(12)] \quad (5\text{-}8)$$

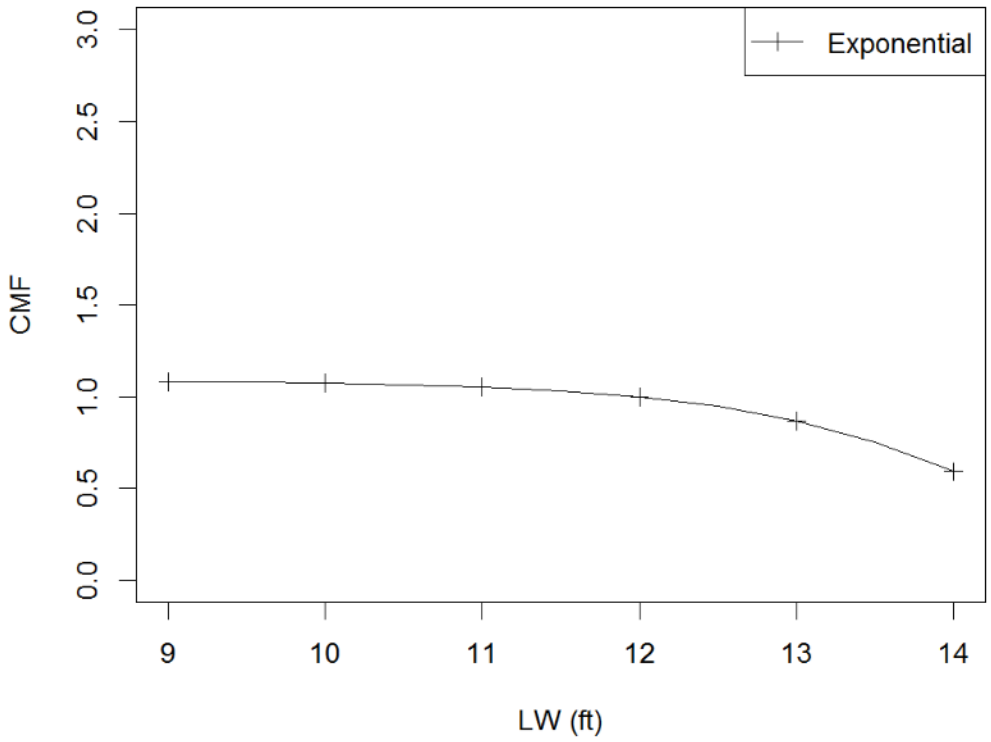


Figure 19 CM-Function for Lane Width Derived using Observed Data (Exponential)

5.2.4 Log

The log form is shown in Equation 5-9.

$$U(LW) = \ln(CR) = A + B \times ln(LW) \quad (5\text{-}9)$$

The results are shown in Tables 37 and 38.

Table 37 Fitting Result of Log Functional Form for Lane Width

Model Variable	Coef. Value [a]	SE [b]	p-Value
A	0.0242	0.0050	< 2e-16
B	-0.0088	0.0020	1.48E-05

Note: a – estimated coefficient value; b – SE = standard error.

Table 38 Modeling Result of Observed Date with Log Functional Form

Model Variable	Coef. Value [a]	SE [b]	p-Value
Intercept [$ln(\beta_0)$]	-7.17	0.20	< 2e-16
Ln(AADT) (β_1)	1.02	0.02	< 2e-16
U(LW) (β_2)	214.47	34.12	3.28E-10
ϕ	1.3349	0.0542	-
AIC		22082.6	
MAD		1.210	
MSPE		5.415	

Note: a – estimated coefficient value; b – SE = standard error.

The CM-Function for lane width is shown Equation 5-10. And the curve is illustrated in Figure 20.

$$ln(CMF) = \beta_2 \times U(LW) - \beta_2 \times U(12) = 0.71 \times [exp(LW) - exp(12)] \qquad (5\text{-}10)$$

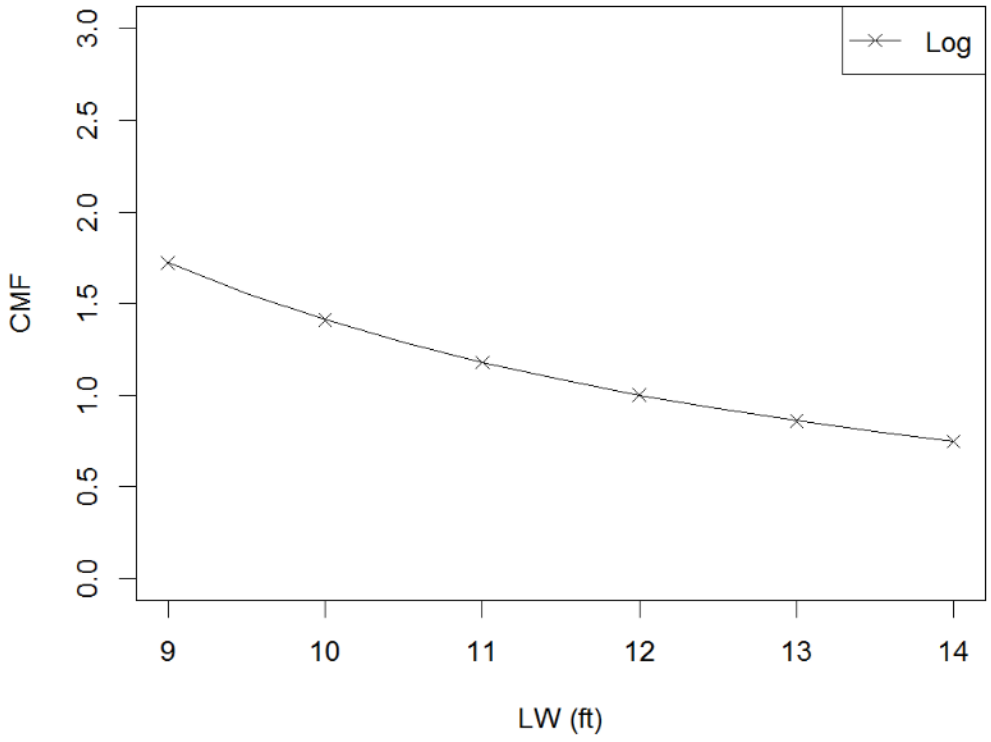


Figure 20 CM-Function for Lane Width Derived using Observed Data (Log)

5.2.5 Power

The power form is shown in Equation 5-11.

$$U(LW) = \ln(CR) = A + LW^{B} \quad (5\text{-}11)$$

The results are shown in Tables 39 and 40.

Table 39 Fitting Result of Power Functional Form for Lane Width

Model Variable	Coef. Value [a]	SE [b]	p-Value
A	-0.9755	0.0051	< 2e-16
B	-0.0090	0.0021	2.25E-05

Note: a – estimated coefficient value; b – SE = standard error.

Table 40 Modeling Result of Observed Date with Power Functional Form

Model Variable	Coef. Value [a]	SE [b]	p-Value
Intercept [$ln(\beta_0)$]	-7.17	0.20	< 2e-16
Ln(AADT) (β_1)	1.02	0.02	< 2e-16
U(LW) (β_2)	214.40	34.11	3.26E-10
ϕ	1.3349	0.0542	-
AIC		22082.6	
MAD		1.210	
MSPE		5.415	

Note: a – estimated coefficient value; b – SE = standard error.

The CM-Function for lane width is shown Equation 5-12. And the curve is illustrated in Figure 21.

$$ln(CMF) = \beta_2 \times U(LW) - \beta_2 \times U(12) = 214.4 \times (LW^{0.009} - 12^{0.009}) \quad (5\text{-}12)$$

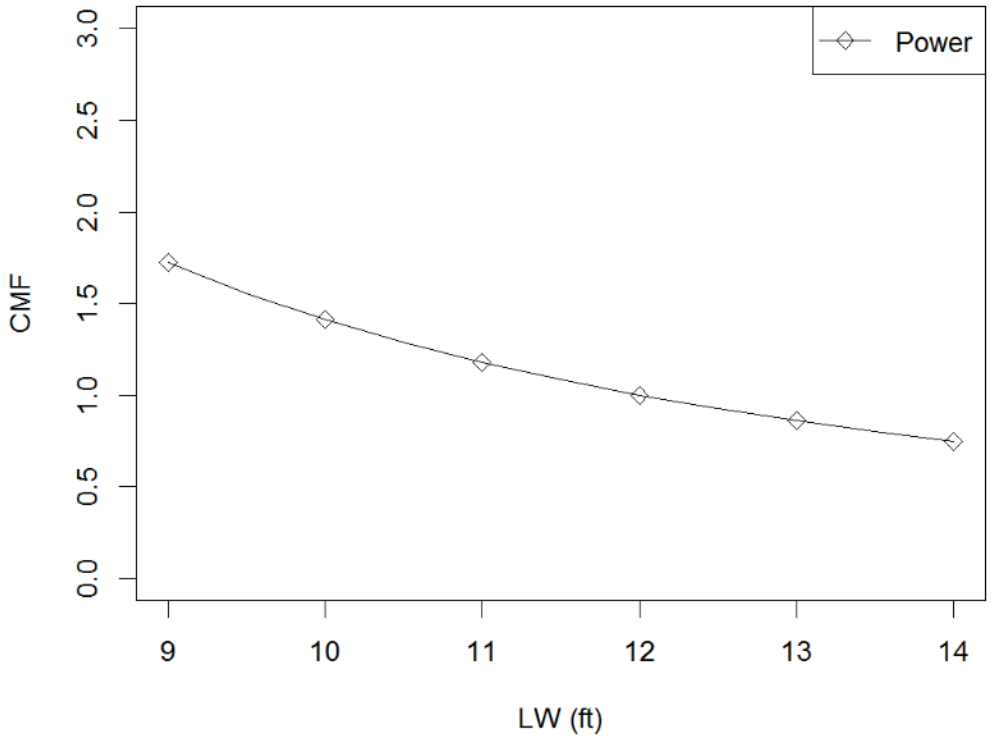


Figure 21 CM-Function for Lane Width Derived using Observed Data (Power)

5.2.6 Quadratic

The quadratic form is shown in Equation 5-13.

$$U(LW) = \ln(CR) = A + B \times LW + C \times LW^2 \quad (5\text{-}13)$$

The results are shown in Tables 41 and 42.

Table 41 Fitting Result of Quadratic Functional Form for Lane Width

Model Variable	Coef. Value [a]	SE [b]	p-Value
A	0.0471	0.0194	1.50E-02
B	-0.0069	0.0033	3.73E-02
C	0.0003	0.0001	6.27E-02

Note: a – estimated coefficient value; b – SE = standard error.

Table 42 Modeling Result of Observed Date with Quadratic Functional Form

Model Variable	Coef. Value [a]	SE [b]	p-Value
Intercept [$ln(\beta_0)$]	-7.15	0.19	< 2e-16
Ln(AADT) (β_1)	1.02	0.02	< 2e-16
U(LW) (β_2)	204.74	30.31	1.42E-11
ϕ	1.3379	0.0545	-
AIC		22078.3	
MAD		1.208	
MSPE		5.376	

Note: a – estimated coefficient value; b – SE = standard error.

The CM-Function for lane width is shown Equation 5-14. And the curve is illustrated in Figure 22.

$$ln(CMF) = \beta_2 \times U(LW) - \beta_2 \times U(12) = 0.05 \times \mathrm{LW}^2 - 1.41 \times \mathrm{LW} + 9.2 \qquad (5\text{-}14)$$

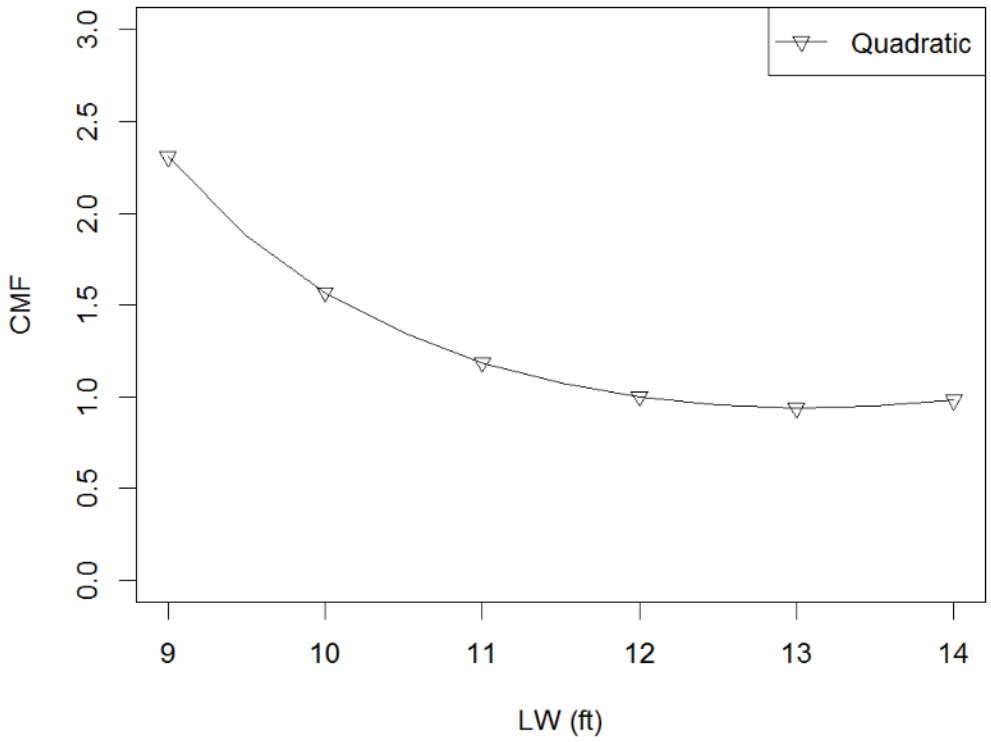


Figure 22 CM-Function for Lane Width Derived using Observed Data (Quadratic)

5.2.7 Comparison of CMFs with various forms

The previous sections provide the CM-Functions for lane width derived from regression models with six types of link functions. They are shown together in Figure 23. It can be seen from Figure 23 that the overall trend of these CM-Functions were the same, widening lane could bring safety benefits. However, there were significant

differences between the six CM-Functions (except that the log and power forms were nearly identical). The inverse, log, and power CM-Functions were close to the linear one. Widening the lane by one feet had fixed (or approximately fixed in the three nonlinear forms) safety effect regardless of the initial width, and this effect was about 15 percent reduction of crashes. But the other two (exponential and quadratic) showed obvious difference. In the exponential CM-Function, when the lane width was between 9 ft and about 12 ft, the CMF was consistently about 1.0, meaning changing the lane width in this range had little effects on safety. But when the lane was wider than 12 ft, widening it reduced crashes significant. In addition, the wider the initial lane was, the more safety benefits widening it could bring (i.e., widening from 13 ft to 14 ft had more effects than that from 12 ft to 13 ft). The quadratic CM-Functions illustrated a different effect. When the lane was narrow (between 9 ft and 12 ft), widening it had significant safety effect. And the narrower the initial lane was, the more effects widening by one foot had. But when the lane was wide (more than 12 ft), widening the lane had relatively minor effect. Specifically, from 12 ft to 13 ft the expected crashes reduced by about six percent, and from 13 ft to 14 ft, it increased by about five percent.

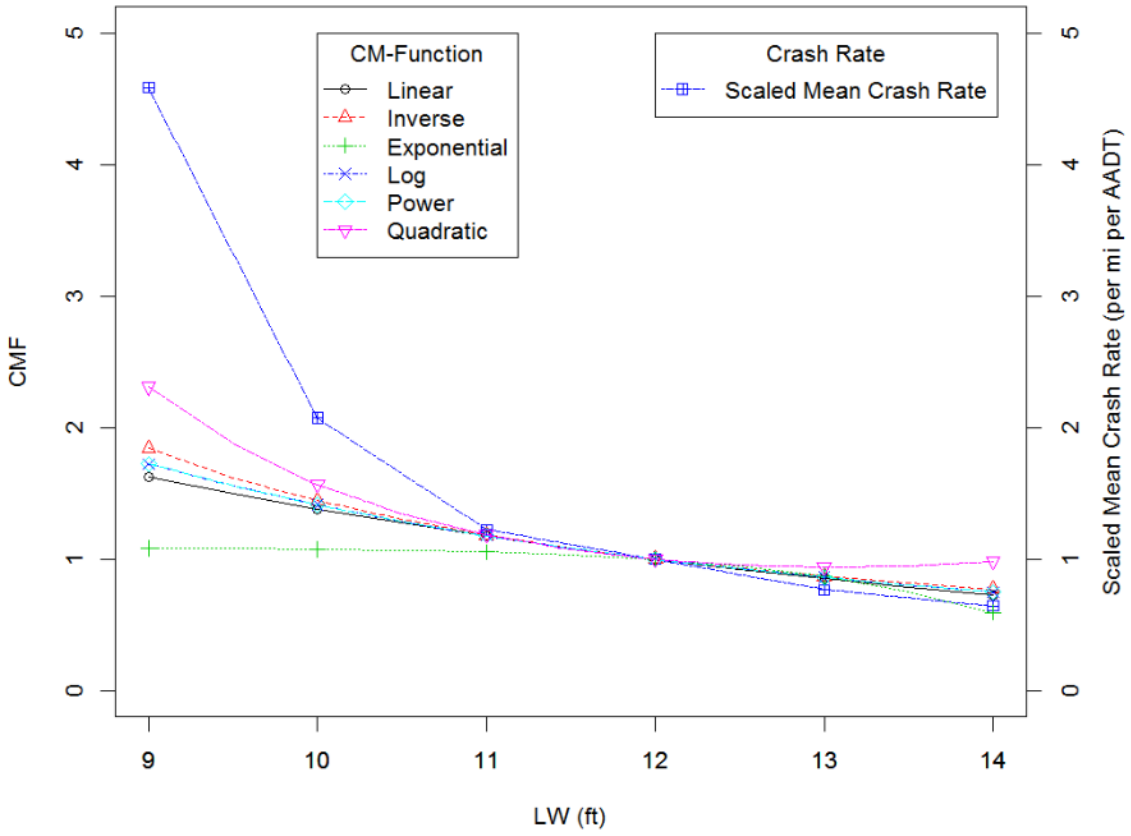


Figure 23 CM-Functions for Lane Width Derived using Observed Data (All)

Since the true safety effect of lane width was unknown, it was difficult to evaluate the quality of the six CM-Functions directly. The GOF and prediction measurements of the six models are shown in Table 43. It can been seen that quadratic function consistently had the lowest AIC, MAD and MSPE. From this perspective, the regression model with quadratic link functions fitted the data best. If a bold assumption, that the mean crash rate (mean number of crashes per mile per AADT) could represent the actual effect of lane width on safety, a "true" CM-Function could be obtained by scaling the mean crash rate (it was scaled such that the rate equaled to 1.0 at base

condition, 12-ft lane.) The curve of the "true" CM-Function is also illustrated in Figure 23. The area between the "true" curve and the six CM-Function curves were calculated, respectively, as shown in Table 43 (row of "Area"). It can be seen that quadratic form had the smallest area with the curve of "true" CM-Function, which made it overall the closest to the "true" safety effect.

Table 43 Comparison of GOFs and Prediction Measurements

Form	Linear	Inverse	Exponential	Log	Power	Quadratic
AIC	22083.91	22081.46	22107.40	22082.61	22082.60	**22078.34**
MAD	1.2106	1.2101	1.2139	1.2103	1.2103	**1.2082**
MSPE	5.4179	5.4111	5.4149	5.4148	5.4148	**5.3758**
Area *	2.353	2.233	3.067	2.311	2.310	**2.014**

Note: the number in bold indicate the smallest in the corresponding row, and those with underline mean they are the largest. * Area = the area between the curve of scaled mean crash rate and that of corresponding CM-Function.

Given the "true" CMFs for lane width, the "bias" and "error percentage" of those derived from regression models with various function forms were calculated (using the same method described in Section 3.3.2) at several points of interest, the results are shown in Table 44. First, none of the six CM-Functions adequately captured the "true" safety effect of lane width. They all underestimated the CMFs on the left side (i.e., lane narrower than 12 ft) and vice versa on the right side (lane wider than 12 ft). Second, none could always outperform others. For example, quadratic form had the smallest

error percentage when lane width was less than 12 ft, but it had the greatest when lane width was more than 12 ft. Although quadratic form fitted the data best and it was the closest to the "true" curve, one still could not conclude that it was the best. Finally, recall that all the parameters estimated in the six models were statistically significant with at least a 90 percent level. So, even though the modeling result looked statistically correct and acceptable, the CM-Functions derived from the models might be biased when improper link functional forms were used. All of these were consistent with the simulation findings.

5.3 Volume-Only Model and "Full-Variable" Model

The simulation analyses have demonstrated how omitted variables affected the quality of CMFs. In practice, however, it is nearly impossible to capture all factors affecting crash risk. Section 5.2 has shown that lane width influenced crash counts considerably. To explore how the modeling result changed when some variables affecting safety was omitted, this section further analyzed the volume-only and "full-variable" models. The former one, including AADT as the only explanatory variable as the name implies, has been developed often in highway safety.

5.3.1 Volume-Only Model

The target functional form in the volume-only model is shown in Equation 5-15. And the modeling results are presented in Table 45.

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \tag{5-15}$$

Table 44 "Bias" and "Error" of CMFs for Lane Width

LW (ft)	9			10			11			13			14		
"True" CMF	4.59			2.07			1.22			0.77			0.64		
Function Form	CMF	Bias	E*	CMF	Bias	E*	CMF	Bias	E*	CMF	Bias	E*	CMF	Bias	E*
Linear	1.62	2.96	64.6	1.38	0.69	33.4	1.17	0.05	3.9	0.85	-0.09	**11.2**	0.72	-0.08	12.6
Inverse	1.84	2.74	59.8	1.44	0.63	30.4	1.18	0.04	3.4	0.87	-0.10	13.4	0.77	-0.13	19.6
Exponential	1.07	3.51	76.6	1.06	1.01	48.6	1.05	0.18	14.3	0.88	-0.12	15.2	0.63	0.02	**2.4**
Log	1.71	2.87	62.7	1.41	0.67	32.2	1.18	0.05	3.8	0.86	-0.10	12.5	0.75	-0.11	16.5
Power	1.71	2.87	62.7	1.41	0.67	32.2	1.18	0.05	3.8	0.86	-0.10	12.5	0.75	-0.11	16.5
Quadratic	2.38	2.21	**48.1**	1.59	0.48	**23.2**	1.19	0.03	**2.4**	0.94	-0.17	22.4	0.98	-0.34	52.6

Note: the number in bold indicate the smallest in the corresponding column, and those with underline mean they are the largest. * E = error percentage, %.

Table 45 Modeling Result of Volume-Only Model

Model Variable	Coef. Value [a]	SE [b]	p-Value
Intercept [$ln(\beta_0)$]	-6.40	0.15	<2E-16
Ln(AADT) (β_1)	1.00	0.02	<2E-16
AIC		22117.87	
MAD		1.212	
MSPE		5.364	

Note: a – estimated coefficient value; b – SE = standard error.

5.3.2 "Full-Variable" Model

In the contrast, full-variable model included all the variables that affecting crash risk, which was nearly unattainable in practice. In this section, "full-variable" refers to including all the variables available (i.e., AADT, lane width, and shoulder width) in the dataset. Based on the simulation findings, the modeling results can be biased if variables already known to have significant effect on safety are omitted in the analysis. To simplify the problem, in this section, it was assumed that factors other than lane width and shoulder width had trivial effects.

The functional form utilized in the "full-variable" model is shown in Equation 5-16. The modeling results are presented in Table 46.

$$E(\Lambda_i) = \beta_0 \times L_i \times AADT^{\beta_1} \times exp[\beta_2 \times U(LW_i) + \beta_3 \times Shoulder] \quad (5\text{-}16)$$

Where,

$U(LW)$ = link function for lane width, the quadratic results was adopted;

Shoulder = shoulder width (ft).

Table 46 Modeling Result of "Full-Variable" Model

Model Variable	Coef. Value [a]	SE [b]	p-Value
Intercept [$ln(\beta_0)$]	-7.11	0.19	< 2e-16
Ln(AADT) (β_1)	1.05	0.02	< 2e-16
U(LW) (β_2)	172.8	30.7	1.83e-08
Shoulder (β_3)	-0.016	0.0033	9.36e-07
AIC		22058.2	
MAD		1.185	
MSPE		5.052	

Note: a – estimated coefficient value; b – SE = standard error.

5.3.3 Comparison between Volume-Only and "Full-Variable" Models

When comparing the fitting results between the two models, the AIC, MAD and MSPE of the volume-only model were significantly higher than that of the "full-variable" model. This indicates the volume-only model might be relatively less reliable, as expected.

Another significant difference was the coefficient estimates. AADT was the common variable to the two models, but the estimated coefficients for it differed. It was

1.00 and 1.05 in the volume-only model and the "full-variable" model, respectively. Lane width and shoulder width were not included in the volume-only model, and it was impossible to compare the coefficients for these two variables. It can be seen that omitting lane width and shoulder width in the model affected the estimated coefficient of AADT. This is in general consistent with the simulation findings. Since the true relationship between AADT and crash risk was unknown, it was difficult to assess which coefficient represented the relationship better. The lower GOF measurements (i.e., AIC, MAE and MSPE) with the "full-variable" model indicated the estimate within it might be relatively more reliable. In addition, Mallows C_p were calculated for the volume-only model. It was 187.0, which was significantly greater than the number of parameters, 2. This meant the volume-only model was heavily biased.

Besides estimated coefficients, safety practitioners may be more interested in confidence intervals or prediction intervals, because they can provide the uncertainty and are usually more important for use in safety-decision making (Ash et al. 2016). This study calculated the confidence intervals (CIs) for the estimated safety (M) and the prediction intervals (PIs) for the predicted number of crashes (Y) using the two models separately. The approach proposed by Wood (2005), Lord (2008) and Lord et al. (2010) was utilized. Since the two models contained different variables, the CIs and PIs for three kinds of segments were calculated, as shown below. The AADT varied between 500 and 15,000, and the segment length was 0.1 mile.

(a) Normal segment. 12-ft lane and 8-ft shoulder;

(b) Narrow segment: 11-ft lane and 3-ft shoulder; and,

(c) Wide segment: 12-ft lane and 16-ft shoulder.

The estimated safety (M), its 95 percent CIs, and 95 percent PIs of the predicted number of crashes (Y) as functions of AADT for the three kinds of segments are shown in Figure 24, respectively. Note that the lower bound for both estimated safety (M) and predicted number of crashes (Y) were all zero. For the normal segment, in Figure 24 (a), the estimated safety (M) were nearly the same in the two models, and the CIs and PIs were also very close. But for the other two kinds of segments, they differed a lot. For the narrow segment (i.e., Figure 24 (b)), the safety (M) was greater in the "full-variable" model, because according to the modeling result, narrowing lane and/or shoulder increased the predicted number of crashes. However, in the volume-only model, changes in lane width or shoulder width had no influence on the estimated safety (M) or the predicted number of crashes (Y). Both CIs and PIs of "full-variable" model was wider than that of volume-only model, especially when the AADT was high (as shown in the figures on the right side of Figure 24). The situation was contrary for the wide segment (i.e., Figure 24 (c)). Estimated safety (M) was smaller in "full-variable" model, its CIs and PIs were narrower.

Since the exact values of the safety of these segments are unknown for the observed dataset, no comments can be made on which model performs better regarding the CIs or PIs. Nevertheless, adding or omitting variables influenced the estimated safety and predicted number of crashes as well as the confidence/prediction intervals.

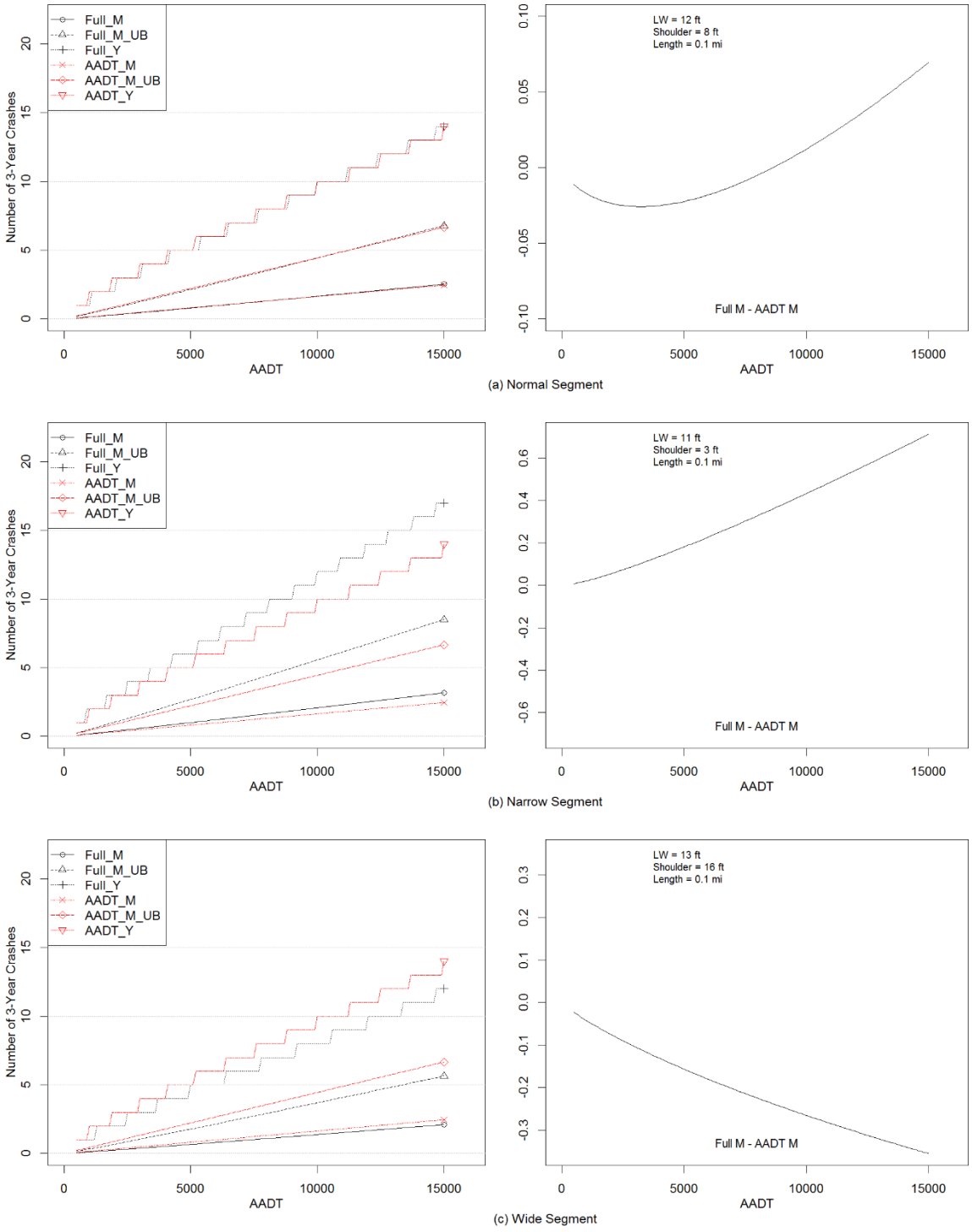


Figure 24 Ninety-Five Percent Confidence Intervals

It is worth to mention that the three variables (i.e., AADT, lane width and shoulder width) were related in the observed dataset, as can be seen in Figure 25. High volume roads were very likely to be wider in lane and somewhat in shoulder, and vice versa. That is to say, even though lane width and shoulder width were omitted in the volume-only model, there were still some information about lane width and shoulder width (in forms of AADT) included in it. In addition, either lane width or shoulder width was uniformly distributed among their range, as shown in Figure 26. This fell under the variable correlation and distribution problems that have been discussed in Section 4.4. This should have also influenced the modeling result.

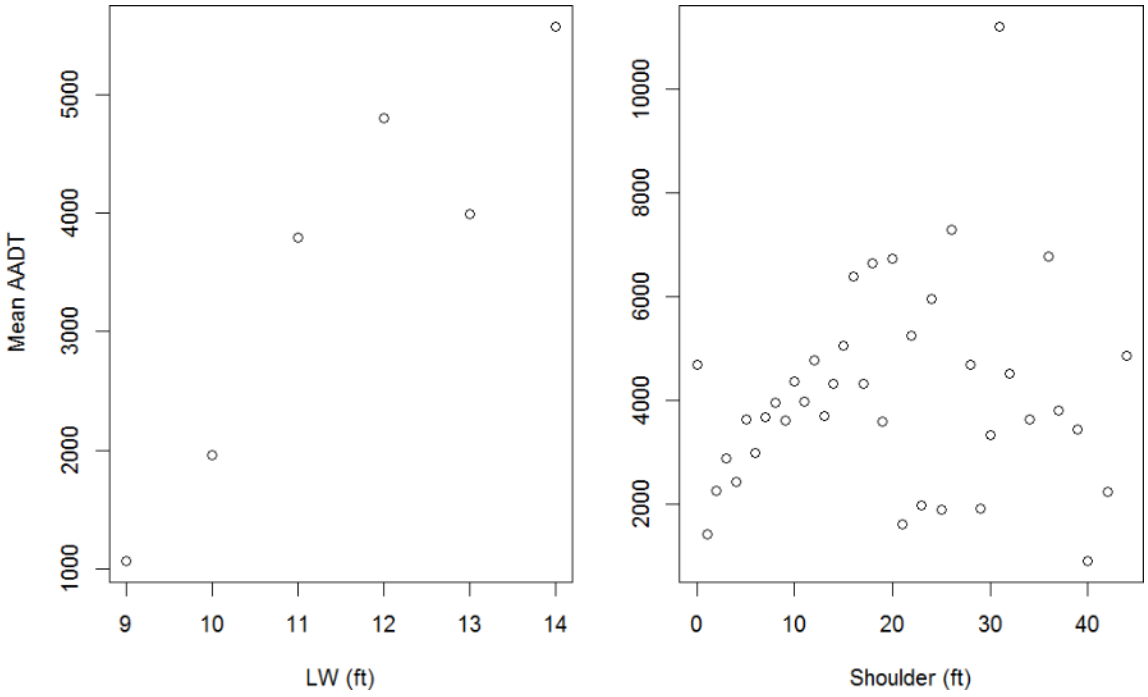


Figure 25 Mean AADT and Lane Width

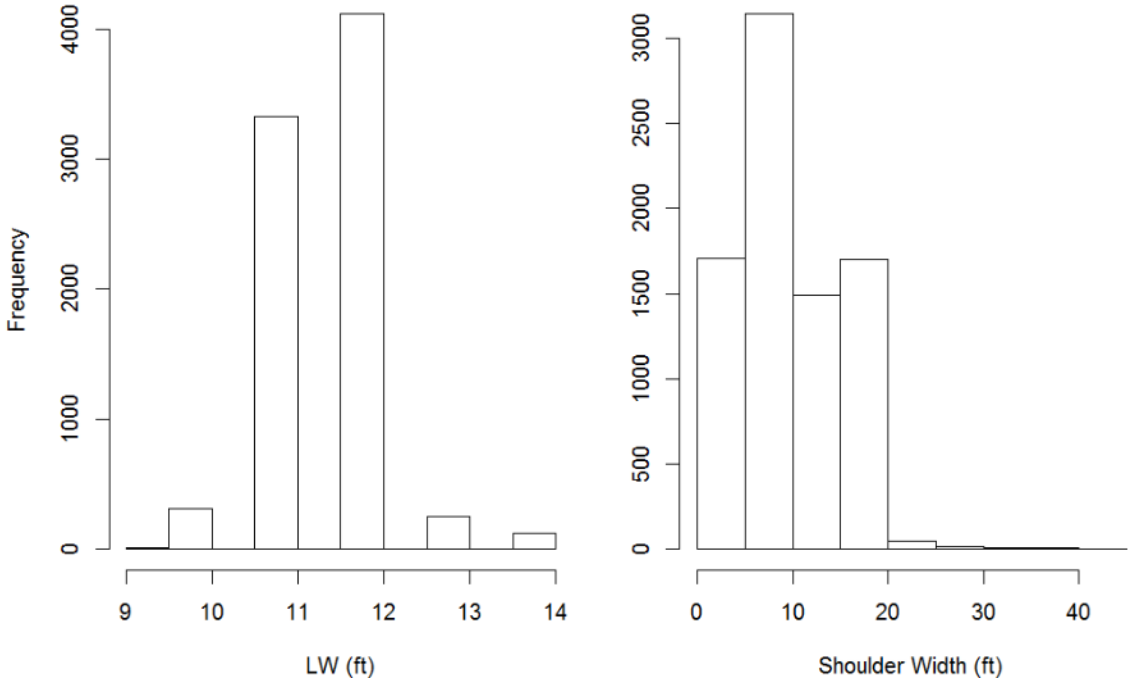


Figure 26 Histograms of Lane Width and Shoulder Width

5.4 Summary

In this chapter, real observed crash data were analyzed to estimate CMFs from various models. The results successfully supported the previous findings based on simulated data. The CM-Functions differed a lot when different functional forms were used in the regression models. Some showed reasonable results, while some looked contrary to the data. In general, the CM-Functions derived from the models with better GOF measures (i.e., AIC, MAE and MSPE) seemed to be more reliable. Omitting variable(s) made the model less reliable, and significantly influenced the estimated coefficients as well as the predicted crash numbers.

The next chapter provides a summary and discussion of the research accomplished in this study, and discusses future research on this topic.

6. SUMMARY AND CONCLUSIONS

The effectiveness evaluation of countermeasures is an important process in roadway safety management. It provides valuable information for future decision making and policy development (AASHTO 2010). CMFs have been widely used to quantify the safety effects of treatments (or changes in design or operation). Before-after study and regression model analysis are the two main approaches used to develop CMFs. The former has been considered to be superior for estimating CMFs in the past two decades or so. However, several limitations have restricted its use for developing high-quality CMFs. As an alternative, regression models have been popular to estimate CMFs in recent years. Nevertheless, both of the two methods have their own drawbacks. So far, no study has comprehensively evaluated whether or not regression models should be used for developing CMFs, because they may not properly capture the safety effect of treatments. Considering the fact that a large number of CMFs have been developed using regression models, it is necessary to evaluate the accuracy of CMFs estimated from this kind of approach. Hence, the primary objective of this study was to examine the use of regression models for developing CMFs. Specifically, the goal was to investigate the accuracy of the CMFs derived from regression models under various conditions. The objective was mainly accomplished using simulated data and the findings were validated using observed data. Several cautions of using regression models for developing CMFs were raised.

The following two sections summarize the work of this study and make some recommendations for developing CMFs using regression models in the future, respectively.

6.1 Summary of This Study

This section briefly summarizes the major contributions of each chapter.

Chapter 1 introduced the general information about developing CMFs and the objective of this study.

In Chapter 2, a background about research findings on CMFs was documented. Specifically, it presented the advantages and limitations of the common CMF estimating approaches (i.e., before-after study and regression model method), the attempts made to explore the nonlinear relationships some highway features had, and the combined safety effects of multiple treatments on safety.

Chapter 3 described detailed methodologies for examining the accuracy of CMFs estimated using regression models. A simulation protocol was developed and an example was provided for illustrating the simulation procedures. Measurement used to quantify the nonlinearity of CM-Function was proposed. And nine scenarios were specified to examine the CMFs under various conditions.

The result of each simulation scenario was discussed in Chapter 4. The main findings are summarized as follows:

(1) The CMFs derived from the common regression models should be unbiased when the premise of cross-sectional studies were met (i.e., all segments were similar, proper functional forms, variables were independent, enough sample size, etc.).

(2) Functional forms played important roles in developing reliable CMFs. Improper functions may lead to misleading conclusions and biased CMFs. This is consistent with several previous studies (Davis 2000, 2014; Hauer 2005a, 2010; Lord and Mannering 2010; Miaou and Lord 2003). When improper forms for some variables were used, the CMFs for these variables were biased, and the quality of CMFs for other variables could also be influenced. Using the common GLMs method to model variables having nonlinear relationships with crash risk would produce biased CMFs. With the increase in nonlinearity, the bias became significant. This might also produce biased estimates for other parameters. In addition, variable correlation and/or distribution showed influence on the CMFs as well as other parameter estimates when improper functional forms were used.

(3) Regression models did suffer from the omitted-variable problem. If some factors having minor safety effects were omitted, the accuracy of estimated CMFs might still be acceptable. However, if some factors already known to have significant effects on crash risk were omitted, the estimated CMFs were generally unreliable.

(4) When the influence on safety of considered variables were not independent, the CMFs produced from the commonly used regression models were biased. The bias was significantly correlated with the adjustment factor (i.e., degree of their dependence). Higher dependence led to significant bias. Under the conditions of dependence, the

coefficients for the variables of interests might be over- or underestimated, and other variables may be under- or overestimated to compensate for the biased estimated coefficients.

Chapter 5 validated the findings in this study using observed data. The results, in general, were consistent with the simulation findings.

6.2 Recommendations and Future Research Area

Although the CMFs derived from regression models should be unbiased when all requirements of cross-sectional studies are met, this assumption can hardly be satisfied when dealing with observed data. Several issues have been raised when using regression models for developing CMFs. When modeling crash data, transportation safety analysts are recommend to answer the following questions: (1) whether variables having significant effects on safety have been omitted; (2) has improper link functions been used in the models; and (3) are the variables dependent when multiple ones are considered?

The first question can be examined through reviewing the segments or sites to detect if they have significant differences in some safety associated factors other than those included in the models. If these factors have significant effects on safety (e.g., relatively large or small CMFs are found from *HSM*, CMF Clearinghouse, and other relevant documents or peer-reviewed papers), the regression models are likely to suffer from the omitted-variable problem. The second one may be analyzed by observing the patterns between variables and crash frequency or crash rate, or by comparing multiple

link functions (Lee et al. 2015; Park and Abdel-Aty 2015b). For the last question, engineering judgment and experiences may need to be considered, because so far with only limited studies on the safety effects of multiple treatments it is not easy to conclude whether or how much two or more safety treatments are dependent of each other.

If any of the above questions exists, the modeling result will probably lead to biased CMFs and misleading conclusions, and attention should be paid.

There are a few limitations with this study. First, a solid model is the base for developing reliable CMFs when using the regression method. This study used the most common one, NB distribution, to analyze the simulated crash counts, which were generated from the same distribution. Although NB distribution is the most popular one used by safety analysts, new models are being proposed for predicting crashes and some show better results given particular characteristics of crash datasets (Geedipally et al. 2012; Lord et al. 2005; Wu et al. 2014; Zou et al. 2013a; Zou et al. 2013b). Second, sample size influences the modeling result significantly (Lord 2006; Lord and Miranda-Moreno 2008; Ye and Lord 2014). The sample size of simulated dataset was 1,492, and 8,132 for real data. Both should be large enough in this study. These two problems will affect the modeling result and the quality of CMFs. To estimate reliable CMFs, these questions need further consideration when dealing with real observed data. Nevertheless, the simulation protocol proposed in this study can still be applied to evaluate the CMFs under different conditions.

In addition, several problems may exist simultaneously in one regression model. For example, a regression analysis using observed data may have omitted important

variables and utilized improper function form, while the variables included in the model may be dependent. This is likely to make the result worse, and the bias of CMFs should become higher. This study did not examine how multiple problems affected the accuracy of CMFs.

Although before-after studies have been considered to be the state-of-the-art method and are always preferred for developing CMFs, recent studies have pointed out that the before-after study can also be biased (Kuo and Lord 2013; Lord and Kuo 2012). Hence, further research is needed to provide guidelines when a cross-sectional study should be used over the before-after study, and vice versa, as a function of the characteristics of the data.

REFERENCES

AASHTO. 2004. *A policy on geometric design of highways and streets, 2004*. Washington, D.C.: Washington, D.C. : American Association of State Highway and Transportation Officials.

AASHTO. 2010. *Highway safety manual*. 1st Edition ed. Washington, D.C.: American Association of State Highway and Transportation Officials.

Ash, J. E., Y. Zou, D. Lord, and Y. Wang. 2016. "Comparison of confidence and prediction intervals for different mixed-poisson regression models." the Transportation Research Board (TRB) 95th Annual Meeting, Washington, D.C.

Bauer, K., and D. Harwood. 2013. "Safety effects of horizontal curve and grade combinations on rural two-lane highways." *Transportation Research Record: Journal of the Transportation Research Board* 2398:37-49.

Bonneson, J., and D. Lord. 2005. "Role and application of accident modification factors in the highway design process." FHWA/TX-05/0-4703-2. Texas Tranportation Institute, College Station, TX

Bonneson, J., and M. P. Pratt. 2010. Highway safety design workshops. Texas: Texas Transportation Institute.

Carter, D., R. Srinivasan, F. Gross, and F. Council. 2012. "Recommended protocols for developing crash modification factors." Accessed June 26, 2014. http://www.cmfclearinghouse.org/collateral/CMF_Protocols.pdf.

Chen, Y., and B. Persaud. 2014. "Methodology to develop crash modification functions for road safety treatments with fully specified and hierarchical models." *Accident Analysis & Prevention* 70:131-139.

CMFClearinghouse. 2014. "Installation of fixed combined speed and red light cameras." Accessed July 23, 2015. http://www.cmfclearinghouse.org/study_detail.cfm?stid=401.

Davis, G. A. 2000. "Accident reduction factors and causal inference in traffic safety studies: A review." *Accident Analysis & Prevention* 32 (1):95-109.

Davis, G. A. 2014. "Crash reconstruction and crash modification factors." *Accident Analysis & Prevention* 62:294-302.

De Pauw, E., S. Daniels, T. Brijs, E. Hermans, and G. Wets. 2014. "To brake or to accelerate? Safety effects of combined speed and red light cameras." *Journal of Safety Research* 50:59-65.

Elvik, R. 2009. "An exploratory analysis of models for estimating the combined effects of road safety measures." *Accident Analysis & Prevention* 41 (4):876-880.

FHWA. 2010. "CMF clearinghouse brochure." Accessed January 20, 2014. http://www.cmfclearinghouse.org/collateral/CMF_brochure.pdf.

FHWA. 2011. "Highway safety information system." Accessed September 6, 2015. http://www.hsisinfo.org/.

Geedipally, S. R., D. Lord, and S. S. Dhavala. 2012. "The negative binomial-lindley generalized linear model: Characteristics and application using crash data." *Accident Analysis & Prevention* 45:258-265.

Gross, F., and E. T. Donnell. 2011. "Case–control and cross-sectional methods for estimating crash modification factors: Comparisons from roadway lighting and lane and shoulder width safety effect studies." *Journal of Safety Research* 42 (2):117-129.

Gross, F., and A. Hamidi. 2011. "Investigation of existing and alternative methods for combining multiple CMFs." Accessed July 14, 2015. http://www.cmfclearinghouse.org/collateral/Combining_Multiple_CMFs_Final.pdf.

Gross, F., A. Hamidi, and K. Yunk. 2012. "Issues related to the combination of multiple CMFs." the 91st Annual Meeting of the Transportation Research Board, Washington D.C.

Gross, F., P. P. Jovanis, K. Eccles, and K.-Y. Chen. 2009. Safety evaluation of lane and shoulder width combinations on rural, two lane, undivided roads. edited by U.S. Department of Transportation FHWA. Washington D.C.

Gross, F., C. Lyon, B. Persaud, and R. Srinivasan. 2013. "Safety effectiveness of converting signalized intersections to roundabouts." *Accident Analysis & Prevention* 50:234-241.

Gross, F., B. Persaud, and C. Lyon. 2010. A guide to developing quality crash modification factors. edited by U.S. Department of Transportation FHWA. Washington, D.C.: FHWA, U.S. Department of Transportation.

Hall, J. W., K. L. Smith, L. Titus-Glover, J. C. Wambold, T. J. Yager, and Z. Rado. 2009. "Guide for pavement friction." Transportation Research Board Accessed November 11, 2014. http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w108.pdf.

Hamner, B. 2013. "Package 'metrics'." Accessed May 22, 2014. http://cran.r-project.org/web/packages/Metrics/Metrics.pdf.

Harkey, D. L., R. Srinivasan, J. Baek, F. M. Council, K. Eccles, N. Lefler, F. Gross, B. Persaud, C. Lyon, E. Hauer, and J. A. Bonneson. 2008. *Accident modification factors for traffic engineering and its improvements*. Edited by David L. Harkey, Program National Cooperative Highway Research, Board National Research Council . Transportation Research and Officials American Association of State Highway and Transportation. Washington, D.C.: Washington, D.C. : Transportation Research Board.

Hauer, E. 1991. "Should stop yield - matters of method in safety research." *ITE Journal-Institute of Transportation Engineers* 61 (9):25-31.

Hauer, E. 1997. *Observational before-after studies in road safety: Estimating the effect of highway and traffic engineering measures on road safety*. Tarrytown, N.Y., U.S.A.: Pergamon.

Hauer, E. 2004. "Statistical road safety modeling." *Transportation Research Record: Journal of the Transportation Research Board* 1897:81-87.

Hauer, E. 2005a. "Cause and effect in observational cross-section studies on road safety." the 84th Annual Meeting of the Transportation Research Board (TRB), Washington D.C.

Hauer, E. 2005b. "Fishing for safety information in murky waters." *Journal of Transportation Engineering* 131 (5):340-344.

Hauer, E. 2010. "Cause, effect and regression in road safety: A case study." *Accident Analysis & Prevention* 42 (4):1128-1135.

Hauer, E. 2014. "Trustworthiness of safety performance functions." the 93rd Annual Meeting of the Transportation Research Board (TRB), Washington, D.C.

Hauer, E. 2015. *The art of regression modeling in road safety*: Springer.

Hauer, E., F. M. Council, and Y. Mohammedshah. 2004. "Safety models for urban four-lane undivided road segments." *Transportation Research Record: Journal of the Transportation Research Board* 1897:96-105.

Hauer, E., D. W. Harwood, F. M. Council, and M. S. Griffith. 2002. "Estimating safety by the empirical bayes method - a tutorial." *Transportation Research Record: Journal of the Transportation Research Board* 1784:126-131.

Hauer, E., and B. Persaud. 1983. "Common bias in before-and-after accident comparisons and its elimination." *Transportation Research Record: Journal of the Transportation Research Board* 905:164-174.

Kuo, P.-F., and D. Lord. 2013. "Accounting for site-selection bias in before-after studies for continuous distributions: Characteristics and application using speed data." *Transportation Research Part A* 49:256-269.

Kutner, M. H., C. Nachtsheim, J. Neter, and W. Li. 2005. *Applied linear statistical models*. Boston: McGraw-Hill Irwin,.

Lao, Y., G. Zhang, Y. Wang, and J. Milton. 2014. "Generalized nonlinear models for rear-end crash risk analysis." *Accident Analysis & Prevention* 62:9-16.

Lee, C., M. Abdel-Aty, J. Park, and J.-H. Wang. 2015. "Development of crash modification factors for changing lane width on roadway segments using generalized nonlinear models." *Accident Analysis & Prevention* 76:83-91.

Li, X., D. Lord, and Y. Zhang. 2011. "Development of accident modification factors for rural frontage road segments in texas using generalized additive models." *Journal of Transportation Engineering* 137 (1):74-83.

Li, X., D. Lord, Y. Zhang, and Y. Xie. 2008. "Predicting motor vehicle crashes using support vector machine models." *Accident Analysis & Prevention* 40 (4):1611-8.

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

Lord, D. 2008. "Methodology for estimating the variance and confidence intervals for the estimate of the product of baseline models and amfs." *Accident Analysis & Prevention* 40 (3):1013-1017.

Lord, D., and J. A. Bonneson. 2007. "Development of accident modification factors for rural frontage road segments in texas." *Transportation Research Record: Journal of the Transportation Research Board* 2023:20-27.

Lord, D., S. D. Guikema, and S. R. Geedipally. 2008. "Application of the conway-maxwell-poisson generalized linear model for analyzing motor vehicle crashes." *Accident Analysis & Prevention* 40 (3):1123-1134.

Lord, D., and P.-F. Kuo. 2012. "Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures." *Accident Analysis & Prevention* 47:52-63.

Lord, D., P.-F. Kuo, and S. Geedipally. 2010. "Comparison of application of product of baseline models and accident-modification factors and models with covariates." *Transportation Research Record: Journal of the Transportation Research Board* 2147:113-122.

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

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

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

Mannering, F. L., and C. R. Bhat. 2014. "Analytic methods in accident research: Methodological frontier and future directions." *Analytic Methods in Accident Research* 1:1-22.

Math is Fun. 2014. "Definition of fun." Accessed November 28, 2015. https://www.mathsisfun.com/definitions/bias.html.

Miaou, S.-P., and D. Lord. 2003. "Modeling traffic crash flow relationships for intersections - dispersion parameter, functional form, and bayes versus empirical bayes methods." *Transportation Research Record: Journal of the Transportation Research Board* 1840:31-40.

Noland, R. B. 2003. "Traffic fatalities and injuries: The effect of changes in infrastructure and other trends." *Accident Analysis & Prevention* 35 (4):599-611.

Park, B.-J., D. Lord, and C. Lee. 2014a. "Finite mixture modeling for vehicle crash data with application to hotspot identification." *Accident Analysis & Prevention* 71:319-326.

Park, J., and M. Abdel-Aty. 2015a. "Assessing the safety effects of multiple roadside treatments using parametric and nonparametric approaches." *Accident Analysis & Prevention* 83:203-213.

Park, J., and M. Abdel-Aty. 2015b. "Development of adjustment functions to assess combined safety effects of multiple treatments on rural two-lane roadways." *Accident Analysis & Prevention* 75:310-319.

Park, J., M. Abdel-Aty, and C. Lee. 2014b. "Exploration and comparison of crash modification factors for multiple treatments on rural multilane roadways." *Accident Analysis & Prevention* 70:167-177.

Ripley, B., B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, and D. Firth. 2014. "Package 'mass'." Accessed May 20, 2014. http://cran.r-project.org/web/packages/MASS/MASS.pdf.

Roberts, P., and B. Turner. 2007. "Estimating the crash reduction factor from multiple road engineering countermeasures." 3rd International Road Safety Conference, Perth, Australia.

Rodegerdts, L., M. Blogg, E. Wemple, E. Myers, M. Kyte, M. Dixon, G. List, A. Flannery, R. Troutbeck, W. Brilon, N. Wu, B. Persaud, C. Lyon, D. Harkey, and D. Carter. 2007. Roundabouts in the united states. edited by U.S. Department of Transportation. Washington, DC.

Shen, J., and A. Gan. 2003. "Development of crash reduction factors: Methods, problems, and research needs." *Transportation Research Record: Journal of the Transportation Research Board* 1840:50-56.

Srinivasan, R., and D. Carter. 2011. "Development of safety performance functions for north carolina." FHWA/NC/2010-09. University of North Carolina, Chapel Hill, NC

Tarko, A. P., S. Eranky, K. C. Sinha, and R. Scinteie. 1999. "An attempt to develop crash reduction factors using regression technique." the 78th Annual Meeting of Transportation Research Board, Washington, D.C.

TxDOT. 2014. "Roadway design manual." Accessed August 3, 2014. http://onlinemanuals.txdot.gov/txdotmanuals/rdw/rdw.pdf.

Wang, X., T. Wang, A. Tarko, and P. J. Tremont. 2015. "The influence of combined alignments on lateral acceleration on mountainous freeways: A driving simulator study." *Accident Analysis & Prevention* 76:110-117.

Wood, G. 2005. "Confidence and prediction intervals for generalised linear accident models." *Accident Analysis & Prevention* 37 (2):267-273.

Wu, L., D. Lord, and Y. Zou. 2015. "Validation of crash modification factors derived from cross-sectional studies with regression models." *Transportation Research Record: Journal of the Transportation Research Board* 2514:88-96.

Wu, L., Y. Zou, and D. Lord. 2014. "Comparison of sichel and negative binomial models in hot spot identification." *Transportation Research Record: Journal of the Transportation Research Board* 2460:107-116.

Xie, Y., D. Lord, and Y. Zhang. 2007. "Predicting motor vehicle collisions using bayesian neural network models: An empirical analysis." *Accident Analysis & Prevention* 39 (5):922-933.

Xie, Y., and Y. Zhang. 2008. "Crash frequency analysis with generalized additive models." *Transportation Research Record: Journal of the Transportation Research Board* 2061:39-45.

Ye, F., and D. Lord. 2014. "Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models." *Analytic Methods in Accident Research* 1:72-85.

Zou, Y., D. Lord, Y. Zhang, and Y. Peng. 2013a. "Comparison of sichel and negative binomial models in estimating empirical bayes estimates." *Transportation Research Record: Journal of the Transportation Research Board* 2392:11-21.

Zou, Y., L. Wu, and D. Lord. 2015. "Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in negative binomial models." *Analytic Methods in Accident Research* 5–6:1-16.

Zou, Y., Y. Zhang, and D. Lord. 2013b. "Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis." *Accident Analysis & Prevention* 50:1042-1051.

Table 47 Results of CMFs in Scenario IX (ϕ = 1.0)

#	AF [a]	LW [b]				SW [b]				ϕ [c]	AIC[d]	MAD[e]	MSPE[f]
		Th.	SPF (SD)	Bias	E	Th.	SPF (SD)	Bias	E				
IX-1	0.8	0.8	0.72 (0.05)	-0.08	9.88	0.85	0.77 (0.054)	-0.08	8.92	1.00	8901.22	0.057	0.016
IX-2	0.9	0.8	0.74 (0.064)	-0.06	6.98	0.85	0.81 (0.065)	-0.04	4.16	0.97	9040.94	0.056	0.017
IX-3	0.95	0.8	0.79 (0.062)	-0.01	1.16	0.85	0.84 (0.043)	-0.01	1.07	0.99	9019.17	0.044	0.011
IX-4	1.05	0.8	0.81 (0.052)	0.01	1.57	0.85	0.88 (0.077)	0.03	3.21	1.02	9120.98	0.049	0.013
IX-5	1.1	0.8	0.83 (0.064)	0.03	4.02	0.85	0.9 (0.067)	0.05	5.57	1.00	9179.48	0.048	0.013
IX-6	1.2	0.8	0.9 (0.062)	0.10	11.90	0.85	0.92 (0.068)	0.07	8.15	0.98	9334.57	0.059	0.019
IX-7	0.8	0.9	0.82 (0.061)	-0.08	8.62	0.85	0.77 (0.051)	-0.08	9.86	0.99	9173.11	0.063	0.020
IX-8	0.9	0.9	0.85 (0.057)	-0.05	6.01	0.85	0.81 (0.057)	-0.04	4.46	1.01	9240.26	0.048	0.013

Table 47 Continued

#	AF [a]	LW [b]				SW [b]				ϕ [c]	AIC[d]	MAD[e]	MSPE[f]
		Th.	SPF (SD)	Bias	E	Th.	SPF (SD)	Bias	E				
IX-9	0.95	0.9	0.89 (0.06)	-0.01	1.11	0.85	0.84 (0.063)	-0.01	1.24	0.99	9311.30	0.048	0.012
IX-10	1.05	0.9	0.92 (0.046)	0.02	2.35	0.85	0.87 (0.069)	0.02	2.16	0.98	9423.14	0.051	0.015
IX-11	1.1	0.9	0.95 (0.066)	0.05	5.59	0.85	0.9 (0.067)	0.05	5.70	0.99	9462.37	0.053	0.016
IX-12	1.2	0.9	0.98 (0.072)	0.08	8.98	0.85	0.93 (0.059)	0.08	9.48	0.99	9560.22	0.062	0.020
IX-13	0.8	0.8	0.72 (0.05)	-0.08	10.16	0.9	0.82 (0.058)	-0.08	8.92	1.00	9001.58	0.057	0.017
IX-14	0.9	0.8	0.77 (0.054)	-0.03	3.83	0.9	0.85 (0.047)	-0.05	5.18	0.99	9131.19	0.050	0.014
IX-15	0.95	0.8	0.78 (0.049)	-0.02	1.91	0.9	0.89 (0.081)	-0.01	1.60	1.00	9163.99	0.050	0.013
IX-16	1.05	0.8	0.82 (0.053)	0.02	2.78	0.9	0.92 (0.065)	0.02	2.53	1.00	9276.12	0.050	0.013

Table 47 Continued

#	AF [a]	LW [b]				SW [b]				ϕ [c]	AIC[d]	MAD[e]	MSPE[f]
		Th.	SPF (SD)	Bias	E	Th.	SPF (SD)	Bias	E				
IX-17	1.1	0.8	0.83 (0.051)	0.03	4.36	0.9	0.93 (0.064)	0.03	3.84	1.00	9323.39	0.049	0.014
IX-18	1.2	0.8	0.88 (0.057)	0.08	9.59	0.9	0.99 (0.068)	0.09	10.14	1.00	9444.75	0.057	0.017
IX-19	0.8	0.9	0.81 (0.056)	-0.09	9.64	0.9	0.81 (0.064)	-0.09	9.49	1.01	9233.06	0.063	0.020
IX-20	0.9	0.9	0.85 (0.067)	-0.05	5.17	0.9	0.86 (0.075)	-0.04	4.19	0.98	9380.50	0.055	0.015
IX-21	0.95	0.9	0.88 (0.081)	-0.02	1.72	0.9	0.88 (0.064)	-0.02	2.12	1.01	9439.87	0.056	0.017
IX-22	1.05	0.9	0.91 (0.062)	0.01	1.13	0.9	0.93 (0.06)	0.03	3.41	1.00	9561.76	0.050	0.013
IX-23	1.1	0.9	0.96 (0.062)	0.06	6.18	0.9	0.95 (0.064)	0.05	5.88	0.98	9609.89	0.052	0.015
IX-24	1.2	0.9	0.99 (0.071)	0.09	10.00	0.9	0.98 (0.072)	0.08	9.18	0.99	9704.08	0.064	0.022

Note: the same notes as those in Table 29.

Table 48 Results of CMFs in Scenario IX (ϕ = 2.0)

#	AF [a]	LW [b]				SW [b]				ϕ [c]	AIC[d]	MAD[e]	MSPE[f]
		Th.	SPF (SD)	Bias	E	Th.	SPF (SD)	Bias	E				
IX-1	0.8	0.8	0.73 (0.055)	-0.07	8.60	0.85	0.76 (0.062)	-0.09	10.2	1.99	8712.43	0.065	0.023
IX-2	0.9	0.8	0.76 (0.07)	-0.04	4.87	0.85	0.81 (0.078)	-0.04	4.53	1.99	8852.32	0.062	0.021
IX-3	0.95	0.8	0.76 (0.079)	-0.04	4.75	0.85	0.83 (0.08)	-0.02	1.89	1.99	8859.33	0.064	0.026
IX-4	1.05	0.8	0.82 (0.091)	0.02	2.09	0.85	0.88 (0.087)	0.03	3.28	2.02	8969.92	0.064	0.021
IX-5	1.1	0.8	0.84 (0.066)	0.04	4.41	0.85	0.9 (0.073)	0.05	5.86	1.99	9019.91	0.065	0.031
IX-6	1.2	0.8	0.87 (0.079)	0.07	9.14	0.85	0.92 (0.074)	0.07	8.10	2.01	9145.89	0.071	0.028
IX-7	0.8	0.9	0.81 (0.081)	-0.09	10.47	0.85	0.78 (0.075)	-0.07	8.50	1.98	8937.33	0.074	0.027
IX-8	0.9	0.9	0.86 (0.081)	-0.04	4.89	0.85	0.81 (0.067)	-0.04	4.32	2.00	9092.99	0.063	0.021

Table 48 Continued

#	AF [a]	LW [b]				SW [b]				ϕ [c]	AIC[d]	MAD[e]	MSPE[f]
		Th.	SPF (SD)	Bias	E	Th.	SPF (SD)	Bias	E				
IX-9	0.95	0.9	0.87 (0.07)	-0.03	3.20	0.85	0.82 (0.069)	-0.03	3.23	1.99	9139.26	0.059	0.018
IX-10	1.05	0.9	0.92 (0.082)	0.02	2.62	0.85	0.89 (0.079)	0.04	5.25	1.97	9313.70	0.061	0.021
IX-11	1.1	0.9	0.93 (0.086)	0.03	3.69	0.85	0.89 (0.083)	0.04	4.91	1.98	9289.12	0.064	0.022
IX-12	1.2	0.9	0.99 (0.098)	0.09	9.95	0.85	0.93 (0.086)	0.08	9.26	1.96	9363.04	0.075	0.026
IX-13	0.8	0.8	0.73 (0.062)	-0.07	8.46	0.9	0.8 (0.067)	-0.10	11.18	2.02	8823.29	0.069	0.026
IX-14	0.9	0.8	0.76 (0.061)	-0.04	4.38	0.9	0.87 (0.066)	-0.03	3.09	1.98	8995.65	0.065	0.025
IX-15	0.95	0.8	0.79 (0.072)	-0.01	1.40	0.9	0.9 (0.07)	0.00	0.44	2.05	8924.75	0.061	0.020
IX-16	1.05	0.8	0.82 (0.079)	0.02	3.09	0.9	0.92 (0.095)	0.02	2.64	1.98	9053.82	0.062	0.019

Table 48 Continued

#	AF [a]	LW [b]				SW [b]				ϕ [c]	AIC[d]	MAD[e]	MSPE[f]
		Th.	SPF (SD)	Bias	E	Th.	SPF (SD)	Bias	E				
IX-17	1.1	0.8	0.84 (0.067)	0.04	4.74	0.9	0.96 (0.095)	0.06	6.11	2.04	9088.58	0.066	0.023
IX-18	1.2	0.8	0.87 (0.074)	0.07	8.97	0.9	0.98 (0.064)	0.08	8.49	2.00	9237.24	0.067	0.028
IX-19	0.8	0.9	0.82 (0.08)	-0.08	9.24	0.9	0.82 (0.061)	-0.08	8.51	2.01	9045.91	0.073	0.027
IX-20	0.9	0.9	0.86 (0.068)	-0.04	3.90	0.9	0.85 (0.067)	-0.05	5.10	1.99	9248.58	0.062	0.021
IX-21	0.95	0.9	0.87 (0.067)	-0.03	2.93	0.9	0.88 (0.093)	-0.02	2.31	2.01	9210.21	0.066	0.026
IX-22	1.05	0.9	0.89 (0.081)	-0.01	0.82	0.9	0.93 (0.086)	0.03	3.66	2.00	9386.44	0.071	0.027
IX-23	1.1	0.9	0.96 (0.084)	0.06	6.62	0.9	0.94 (0.093)	0.04	4.47	1.97	9403.88	0.069	0.026
IX-24	1.2	0.9	0.99 (0.078)	0.09	10.46	0.9	0.98 (0.098)	0.08	8.87	1.99	9509.79	0.078	0.034

Note: the same notes as those in Table 29.