Validation of CMFs Derived from Cross-Sectional Studies

Using Regression Models

By

Lingtao Wu*
Ph.D. Candidate, Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, Texas 77843-3136
Phone: 979-587-3518, fax: 979-845-6481
Email: wulingtao@gmail.com

Dominique Lord, Ph.D.
Associate Professor and Zachry Development Professor I
Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, Texas 77843-3136
Phone: 979-458-3949, fax: 979-845-6481
Email: d-lord@tamu.edu

Yajie Zou, Ph.D.
Research Associate
Department of Civil and Environmental Engineering
133B More Hall, Box 352700
University of Washington, Seattle, Washington 98195-2700
Phone: 936-245-5628, fax: 206-543-5965
Email: zouyajie@gmail.com

Word count: 6,934 Words (5,434 Text + 6 Tables) + 34 References

November, 2014

*Corresponding author
ABSTRACT

Crash modification factors (CMFs) can be used to capture the safety effects of countermeasures and play significant roles in traffic safety management. As an alternative to the before-after study, the regression model method has been widely used for estimating CMFs. Although before-after studies are considered to be superior, the use of regression models for estimating CMFs has never been fully investigated. This paper consequently sought to examine the conditions in which regression models could be used for such purpose. CMFs for three variables, lane width, curve density and pavement friction, were assumed and used for generating random crash counts. Then, CMFs were derived from regression models using the simulated crash data for three different scenarios. The results were then compared with the assumed true value. The study results showed that (1) the CMFs derived from the regression models should be unbiased when all factors affecting traffic safety are identical in all segments, except those of interest; (2) if some factors having minor safety effects are omitted from the models, the accuracy of estimated CMFs can still be acceptable; (3) if some factors already known to have significant effects on crash risk are omitted, the CMFs derived from the regression models are generally unreliable. Thus, depending on the missing variables that are not included in the model, the transportation safety analyst can decide if the CMFs developed from the regression models should be used for highway safety applications.
INTRODUCTION

A crash modification factor (CMF) is a multiplicative factor that can be used to 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 (1; 2). CMFs play significant roles 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 before-after (e.g., naïve or simple before-after, before-after with comparison group and empirical Bayes (EB) before-after), cross-sectional (e.g., regression models and case-control), expert panel studies etc. (2). Among these methods, before-after and cross-sectional studies are the most popular approaches (3). The 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 (2; 4). 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 (5). 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 (5-7). Lord and Kuo (7) 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 described above, researchers have proposed that cross-sectional studies could be used for developing CMFs (8-10). Although different types of cross-sectional studies have been proposed over the years, the regression model remains the method of choice for estimating CMFs, as reflected by the large number of CMFs documented in the Highway Safety Manual (HSM) (11) and FHWA CMF Clearinghouse (1).

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 relationship between crashes and the variables influencing safety (12). In this context, so far, nobody has examined the statistical performance of CMFs that are developed from regression models. Thus, the primary objective of this study is to comprehensively investigate the robustness and accuracy of the CMFs derived from regression models. A secondary objective is to describe the conditions when the CMFs developed from regression models become unreliable and potentially biased. The study objectives were accomplished using simulated data for three different scenarios.
BACKGROUND

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

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 (2; 3). 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 (3; 5). 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. Readers are referred to (2) for more 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 (4).

*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 (5; 13).

*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., 5 crashes per year). These entry criteria lead to a truncated distribution, which influences the before-after estimate (7).

*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 (2; 4).

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 (2). 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 safety performance functions (SPFs). The SPFs can be used to quantify the effect of a specific variable on the crash occurrence and CMFs are then derived from the model coefficients (2; 10).

Many models have been proposed to predict safety performance and hence to develop CMFs or crash modification functions (CM-Functions) (14). Although recent studies have introduced some new models for transportation safety analysis (15-18), the Generalized Linear Model (GLM) with a negative binomial (NB) error structure is still the most popular method for
modeling traffic crashes. Although 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 (2). 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 estimate and incorrect CMFs (14).

*Functional form:* Functional form establishes the relationship between expected crashes and explanatory variables and is a critical part of the modeling process. So far, various forms have been used to link crashes to explanatory variables. But the modeling results tend to be inconsistent when using different functional forms (19). Thus, the CMFs derived from regression models can be biased.

There are also several other known issues with the crash modeling that will affect the CMFs derived from regression methods. Readers are referred to Lord and Mannering (14) for 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 (2; 20; 21). The same approach and dataset can also generate different CMFs when using different models (16; 22-24). 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 (7). Before-after studies are less prone to confounding factors compared to cross-sectional studies (25). Furthermore, well-designed observational before-after studies provide advantages over other safety countermeasure evaluation methods (4). CMFs derived from regression models are suggested to be compared with those from before-after studies (2). However, no study has investigated whether or not the CMFs derived from regression models really reflect the true safety effects of corresponding treatments. Considering the fact that a large number of CMFs in the HSM (11) and CMF Clearinghouse (1) were derived from regression models, it is necessary to evaluate the accuracy of CMFs estimated from this kind of approach.

**METHODOLOGY**

To examine if CMFs derived from regression models, defined below as SPFs, really reflect the true safety effects of roadway features or treatments, the researchers derived CMFs using simulated data. Since the exact safety effect of a feature or treatment is hardly known in practice, this makes it extremely difficult to examine the CMFs when observed crash data are used. By analyzing simulated data, one can compare the CMFs estimated from SPFs with the assumed CMF values. The simulation experiment used in this study is described in the following section.
Simulation Protocol

The simulation experiment for evaluating the CMFs derived from SPFs was proposed by Hauer (12). First, CMFs for some highway geometric features are assumed. Then, random crash counts are simulated based on the assigned values of CMFs. Finally, the CMFs are estimated from the simulated crash data and are compared with the true CMFs. The simulation procedures are described in detail below:

**Step 1: Assign Initial Values**

Assume CMFs for highway geometric features of interest. This study will assume an exponential relationship between a highway geometric feature and its CMF. For example, we can assume that the CMF for lane width is $CMF_{LW_{\text{Assumed}}}$, that is the expected crash frequency will be multiplied or divided by $CMF_{LW_{\text{Assumed}}}$ if the lane width increases or decreases by one foot.

**Step 2: Calculate Mean Values**

Calculate the true crash means for each segment using SPFs and assumed CMFs using Equation 1 (11).

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

Where,

- $N_{\text{true},i}$ = true crash mean for roadway segment $i$ for a certain time period (i.e., one year).
- $N_{\text{spf},i}$ = 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 is assumed to be 1.0 for all segments in this study.

The SPF in this study is adopted from the HSM (11) for rural two-lane highways (the same as the data used in this paper), as shown in Equation 2.

$$N_{\text{spf},i} = \text{AADT}_i \times L_i \times 365 \times 10^{-6} \times e^{-0.312} = 2.67 \times 10^{-4} \times L_i \times \text{AADT}_i$$  \tag{2}

Where,

- $\text{AADT}_i =$ average annual daily traffic 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 $i$ is gamma distributed with the dispersion parameter $\alpha$ (the inverse dispersion parameter $\phi = 1/\alpha$) and mean equal to 1 (26):

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

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

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

Where,

$\mu_i =$ Poisson mean for segment $i$ for a certain time period;

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

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

Thus, the simulated crash counts follow Poisson-Gamma or NB distribution with parameters $\phi$ and $\mu_i$. The probability density function (PDF) is given by Equation (4) (26).

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

Where,

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

$\mu_i =$ the crash mean during a period for segment $i$; and,

$\phi =$ inverse dispersion parameter.

Step 4: Estimate CMFs from the Simulated Crash data Using NB Regression Models

As has been documented in the background, many models and functional forms have been proposed to predict crashes. For this study, we selected the most commonly used GLM model and functional form, as shown in Equation 5 (23). Note that a different parameter for describing the mean of the site, $\Lambda_i$, was used for estimating the models (compared to the one used for the simulation, $\mu_i$).

$$E(\Lambda_i) = \beta_0 \times L_i \times \text{AADT}^h \times \exp\left(\sum_{j=2}^{n} \beta_j \times x_j\right)$$

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 goodness-of-fit (GOF) of the models, we used the following three methods: (1) Akaike information criterion (AIC), (2) Mean absolute deviance (MAD), and (3) Mean-squared predictive error (MSPE). For more information about MAD and MSPE, readers are referred to (27).

Once the model is fitted and coefficients are estimated using the simulated crash data, the CM-Function for variable \( j \) can then be derived as (2; 23):

\[
CMF_{x,j} = \exp[\beta_j \times (x - x_{0,j})]
\]

(6)

Where,

\[ \beta_j \] = 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 indicates the CMF derived from the SPF for variable \( j \) is \( CMF_j = \exp(\beta_j) \), meaning the expected crash frequency will be multiplied or divided by \( CMF_j \) if the variable \( j \) increases or decreases 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 NB Model**

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

\[
\Delta_j = CMF_{j, Assumed} - CMF_{j, C-S}
\]

(7)

\[
e_j = 100 \times \frac{|\Delta_j|}{CMF_{j, Assumed}}
\]

(8)

Where,

\[ \Delta_j \] = 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

\[ CMF_{j, C-S} \] = CMF derived from the SPF for variable \( j \).
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 further shows snapshots of the simulated crash counts for \( N \) segments considering a single variable of lane width. In this table, the CMF is assumed to be 0.9, which means the increase of one foot in lane width can decrease the predicted number of crashes by 10 percent (1.0-0.9), with the baseline condition for a lane width equal to 12 ft. So, the CMF specific to a segment \( i \) can be calculated as Equation 9.

**TABLE 1 Example of Simulated Crash Counts for \( N \) Segments (\( \alpha = 2 \) or \( \phi = 0.5 \))**

<table>
<thead>
<tr>
<th>Seg.</th>
<th>Length (mile)</th>
<th>AADT</th>
<th>LW (^a) (ft)</th>
<th>CMFs</th>
<th>( N ) (_{spf} )</th>
<th>( N ) (_{true} )</th>
<th>Yr 1</th>
<th>Yr 2</th>
<th>Yr 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.113</td>
<td>15360</td>
<td>11</td>
<td>1.111</td>
<td>0.464</td>
<td>0.515</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.213</td>
<td>18420</td>
<td>8</td>
<td>1.524</td>
<td>1.048</td>
<td>1.598</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0.125</td>
<td>4260</td>
<td>9</td>
<td>1.372</td>
<td>0.142</td>
<td>0.195</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.161</td>
<td>10600</td>
<td>10</td>
<td>1.235</td>
<td>0.456</td>
<td>0.563</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.196</td>
<td>12560</td>
<td>12</td>
<td>1.000</td>
<td>0.658</td>
<td>0.658</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>N</td>
<td>0.234</td>
<td>4580</td>
<td>13</td>
<td>0.900</td>
<td>0.286</td>
<td>0.258</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\( a \) - LW = Lane Width

The CMF for lane width is given by:

\[
CMF_{LW,i} = 0.9^{LW_{i}-12}
\]  

(9)

Where,

\( LW_{i} \) = lane width of segment \( i \) (ft); and

\( CMF_{LW,i} \) = specific CMF for lane width of segment \( i \).

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

\[
N_{true,i} = N_{spf,i} \times CMF_{LW,i}
\]  

(10)

Then, the \( \exp(\epsilon_{i}) \) of each segment is randomly generated based on a Gamma distribution with parameters mean = 1 and dispersion parameter = \( \alpha \), which has the value of 2 in Table 1. \( \mu_{i} \) of segment \( i \) is then calculated by multiplying \( N_{true,i} \) and \( \exp(\epsilon_{i}) \), as shown in Equation (3b).

After, a sequence of Poisson counts can be generated based on the mean \( \mu \) 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 11.

\[
N_{true,i} = N_{spf,i} \times CMF_{LW,i} = 2.67 \times 10^{-4} \times L_{i} \times AADT_{i} \times 0.9^{LW_{i}-12}
\]  

(11a)
Or equivalently,

\[ N_{true,i} = 9.45 \times 10^{-4} \times L_i \times AADT_i \times \exp(-0.105 \times LW_i) \]  \hspace{1cm} (11b)

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

\[ E(\Lambda_i) = \beta_0 \times L_i \times AADT_i \times \exp(\beta_1 \times LW_i) \]  \hspace{1cm} (12)

The coefficients of Equation 12 are estimated using a NB regression model in MASS package (28) within the software R. The GOF measures are calculated using Metrics package (29). The modeling output is shown in Table 2. The p-values indicate the variables are statistically significant at the 1% level in this example. And, the small MAD and MSPE show the modeling result performs well (given the simulated data).

**TABLE 2 Modeling Output of the Example Data (CMF = 0.9, \( \phi = 0.5 \), NB Model)**

<table>
<thead>
<tr>
<th>Model Variable</th>
<th>Theo. Value (^a)</th>
<th>Coef. Value (^b)</th>
<th>SE (^c)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ([\ln(\beta_0)])</td>
<td>(\ln(9.45 \times 10^{-4}) = -6.96)</td>
<td>-6.810</td>
<td>0.340</td>
<td>5.3911E-89</td>
</tr>
<tr>
<td>Ln(AADT) (\beta_1)</td>
<td>1.00</td>
<td>0.960</td>
<td>0.036</td>
<td>9.996E-161</td>
</tr>
<tr>
<td>Lane Width (\beta_2), ft</td>
<td>-0.105</td>
<td>-0.089</td>
<td>0.012</td>
<td>1.5953E-13</td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) – theoretical value; \(^b\) – estimated coefficient value; \(^c\) – SE = standard error.

Based on the fitting result, the CM-Function for lane width derived from this SPF is

\[ CMF_{LW} = \exp[\beta_2 \times (LW - 12)] = 0.915^{LW - 12} \]  \hspace{1cm} (13)

The value of 12 in Equation 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 is expected to be multiplied by \( e^{\beta_2} \approx 0.915 \).

The bias between the assumed CMF and that from this SPF 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 can be estimated. The mean and standard deviation of CMFs and mean of GOF measures can be calculated. The estimation bias and error percentage are then calculated based upon the mean value of derived CMFs.

For illustration purposes, this example only considers one variable, lane width. Three scenarios were examined in this study 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 simulation settings for these three scenarios are described
below.

Scenario I: Consider one variable, the lane width, only. Different CMF values are used for lane
width and the inverse dispersion parameter $\phi$ equals 0.5, 1.0 and 2.0, respectively.

Scenario II: Consider three variables, lane width, curve density and pavement friction. A fixed
CMF value is assigned for each of the three variables. The inverse dispersion
parameter $\phi$ equals 0.5, 1.0 and 2.0, respectively.

Scenario III: Consider three variables, but only one variable is included in the SPF; this falls
under the “omitted bias variables” (14). The inverse dispersion parameter $\phi$ equals
0.5, 1.0 and 2.0, respectively.

DATA DESCRIPTION

The roadway data used in this study contains 1,492 rural two-lane highway segments in Texas.
The variables include segment length, Annual Average Daily Traffic (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 (30). Generally, higher pavement
friction is linked to safer roads. The segment length and AADT are based on actual values from
the Texas data, while the other three are 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 3.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sample Size</th>
<th>Min.</th>
<th>Max</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (mile)</td>
<td>1492</td>
<td>0.1</td>
<td>6.3</td>
<td>0.55 (0.67)</td>
</tr>
<tr>
<td>AADT</td>
<td>1492</td>
<td>502</td>
<td>24800</td>
<td>6643.9 (3996.4)</td>
</tr>
<tr>
<td>Lane Width (ft)</td>
<td>1492</td>
<td>8.0</td>
<td>13.0</td>
<td>10.47 (1.74)</td>
</tr>
<tr>
<td>CD a (per mile)</td>
<td>1492</td>
<td>0.02</td>
<td>16.0</td>
<td>8.1 (4.66)</td>
</tr>
<tr>
<td>PF b</td>
<td>1492</td>
<td>16.0</td>
<td>47.9</td>
<td>31.9 (9.08)</td>
</tr>
</tbody>
</table>

a – CD = Curve Density; b – PF = Pavement Friction; c - SD = Standard Deviation.

It is important to point out that this study selected three geometric features and the CMFs
are 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.
**RESULTS**

**Scenario I: Consider lane width only**

In this scenario, only the lane width is considered. All other factors affecting crash risk are assumed to be identical among all segments. The assumed CMF for lane width varies 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 14, which is similar to Equation 11, but the coefficient of lane width varies.

\[
N_{\text{true},i} = N_{\text{spf},i} \times \text{CMF}_{\text{LW},i} = 2.67 \times 10^{-4} \times L_i \times AADT_i \times \exp[\beta_{\text{LW}} \times (LW_i - 12)]
\]

(14a)

Or equivalently,

\[
N_{\text{true},i} = \beta_{\text{off}} \times L_i \times AADT_i \times \exp(\beta_{\text{LW}} \times LW_i)
\]

(14b)

Where,

\[\beta_{\text{LW}} = \text{coefficient of lane width, varies between -0.163, -0.105, -0.051, 0 and 0.049, corresponding to the assumed CMFs equal to 0.85, 0.90, 0.95, 1.0 and 1.05, respectively; and,}\]

\[\beta_{\text{off}} = \text{offset coefficient, varies between 0.0019, 0.0010, 0.0005, 0.0003 and 0.0002, corresponding to the assumed CMFs equal to 0.85, 0.90, 0.95, 1.0 and 1.05, respectively.}\]

The CMF for each assumed value is derived from the model using the same procedures illustrated in the simulation example. The considered functional form is provided in Equation 12 (now 15 below).

\[E(\Lambda_i) = \beta_0 \times L_i \times AADT^h \times \exp(\beta_2 \times LW_i)\]

(15)

The fitting results are shown in Table 4. 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 is relatively small under different simulation settings. The estimation bias is less than 0.005 for all scenarios, and the error is within 0.5 percent. The small standard deviation of CMFs (second column in Table 4) also indicates the CMFs derived from the experiments are consistent.

Based on the result of Scenario I, it can be concluded that the CMFs derived from this model can reflect the true safety performance of lane width when considering this variable only. In other words, if a regression model is based on a group of roadway segments that are 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.
Table 4 Results of Scenario I

<table>
<thead>
<tr>
<th>Theo. CMF</th>
<th>CMF (SD)</th>
<th>Bias</th>
<th>$E^c$</th>
<th>AIC $^d$</th>
<th>MAD $^e$</th>
<th>MSPE $^f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>0.849 (0.014)</td>
<td>0.001</td>
<td>0.127</td>
<td>11127.63</td>
<td>0.047</td>
<td>0.013</td>
</tr>
<tr>
<td>0.90</td>
<td>0.901 (0.014)</td>
<td>-0.001</td>
<td>0.107</td>
<td>10681.33</td>
<td>0.042</td>
<td>0.011</td>
</tr>
<tr>
<td>0.95</td>
<td>0.949 (0.015)</td>
<td>0.001</td>
<td>0.114</td>
<td>10277.12</td>
<td>0.041</td>
<td>0.010</td>
</tr>
<tr>
<td>1.00</td>
<td>0.998 (0.015)</td>
<td>0.002</td>
<td>0.170</td>
<td>9901.71</td>
<td>0.037</td>
<td>0.008</td>
</tr>
<tr>
<td>1.05</td>
<td>1.051 (0.018)</td>
<td>-0.001</td>
<td>0.109</td>
<td>9540.71</td>
<td>0.036</td>
<td>0.008</td>
</tr>
<tr>
<td>$\phi = 1.0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>0.849 (0.018)</td>
<td>0.001</td>
<td>0.149</td>
<td>11289.108</td>
<td>0.063</td>
<td>0.026</td>
</tr>
<tr>
<td>0.90</td>
<td>0.902 (0.017)</td>
<td>-0.002</td>
<td>0.246</td>
<td>10800.501</td>
<td>0.053</td>
<td>0.017</td>
</tr>
<tr>
<td>0.95</td>
<td>0.945 (0.017)</td>
<td>0.005</td>
<td>0.498</td>
<td>10390.874</td>
<td>0.050</td>
<td>0.015</td>
</tr>
<tr>
<td>1.00</td>
<td>1.002 (0.024)</td>
<td>-0.002</td>
<td>0.233</td>
<td>9995.650</td>
<td>0.048</td>
<td>0.013</td>
</tr>
<tr>
<td>1.05</td>
<td>1.051 (0.019)</td>
<td>-0.001</td>
<td>0.076</td>
<td>9643.008</td>
<td>0.043</td>
<td>0.012</td>
</tr>
<tr>
<td>$\phi = 2.0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>0.853 (0.022)</td>
<td>-0.003</td>
<td>0.395</td>
<td>11020.611</td>
<td>0.083</td>
<td>0.044</td>
</tr>
<tr>
<td>0.90</td>
<td>0.899 (0.023)</td>
<td>0.001</td>
<td>0.099</td>
<td>10574.470</td>
<td>0.068</td>
<td>0.029</td>
</tr>
<tr>
<td>0.95</td>
<td>0.95 (0.024)</td>
<td>0.000</td>
<td>0.018</td>
<td>10185.386</td>
<td>0.065</td>
<td>0.026</td>
</tr>
<tr>
<td>1.00</td>
<td>0.996 (0.026)</td>
<td>0.004</td>
<td>0.436</td>
<td>9827.826</td>
<td>0.062</td>
<td>0.025</td>
</tr>
<tr>
<td>1.05</td>
<td>1.05 (0.025)</td>
<td>0.000</td>
<td>0.032</td>
<td>9454.310</td>
<td>0.056</td>
<td>0.022</td>
</tr>
</tbody>
</table>

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.

Scenario II: Consider three variables with fixed CMFs

In scenario I, only one variable, the lane width, was considered. Scenario II considers a more practical condition: considering lane width, curve density and pavement friction. In this scenario, we assume each of these three variables will influence crash risk, but they are not identical among all segments. The objective is to examine whether the CMFs of multiple variables derived from SPF are reliable.
CMFs for lane width, curve density and pavement friction are assumed to be fixed, and they are 0.90, 1.072 and 0.973, respectively. For curve density, the CMF = 1.072 means that if the curve density of a segment increases by 1 per mile, the expected crash number will increase by 7.2 percent (1.072 - 1.0). And the baseline for curve density is 0 per mile. So, if the curve density of a segment is 0, the specific CMF for curve density of this segment is 1.0. For the pavement friction, the CMF = 0.973 means that if the pavement friction of a segment increases by 1 unit, the expected crash number will decrease by 2.7 percent (1.0 – 0.973). The baseline is 32.

The theoretical function form of the crash data in scenario II is shown in Equation 16.

\[ N_{true,i} = N_{spf,i} \times CMF_{LW,i} \times CMF_{CD,i} \times CMF_{PF,i} \]
\[ = 2.67 \times 10^{-4} \times L_i \times AADT_i \times 0.9^{LW_i} \times 1.072^{CD_i} \times 0.973^{PF_i} \] (16a)

Or equivalently,

\[ N_{true,i} = 0.0023 \times L_i \times AADT_i \times \exp(-0.105LW_i + 0.070CD_i - 0.027PF_i) \] (16b)

\[ E(\Lambda_\ell) = \beta_0 \times L_i \times AADT_i^{\beta_1} \times \exp(\beta_2 \times LW_i + \beta_3 \times CD_i + \beta_4 \times PF_i) \] (17)

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

Based on this experiment, the CMFs derived from SPFs can reflect the true safety performance when considering multiple variables and assuming other safety factors are identical among all segments.
Table 5 Results of Scenario II

<table>
<thead>
<tr>
<th>Variable</th>
<th>Theo. CMF&lt;sup&gt;a&lt;/sup&gt;</th>
<th>CMF (SD)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Bias</th>
<th>E&lt;sup&gt;c&lt;/sup&gt;</th>
<th>AIC&lt;sup&gt;d&lt;/sup&gt;</th>
<th>MAD&lt;sup&gt;e&lt;/sup&gt;</th>
<th>MSPE&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>LW</td>
<td>0.900</td>
<td>0.900 (0.013)</td>
<td>0.000</td>
<td>0.023&lt;sup&gt;g&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>1.072</td>
<td>1.073 (0.006)</td>
<td>-0.001</td>
<td>0.051</td>
<td>13864.7</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>PF</td>
<td>0.973</td>
<td>0.973 (0.002)</td>
<td>0.000</td>
<td>0.040&lt;sup&gt;g&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\phi = 0.5
\]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Theo. CMF&lt;sup&gt;a&lt;/sup&gt;</th>
<th>CMF (SD)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Bias</th>
<th>E&lt;sup&gt;c&lt;/sup&gt;</th>
<th>AIC&lt;sup&gt;d&lt;/sup&gt;</th>
<th>MAD&lt;sup&gt;e&lt;/sup&gt;</th>
<th>MSPE&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>LW</td>
<td>0.900</td>
<td>0.897 (0.014)</td>
<td>0.003</td>
<td>0.366</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>1.072</td>
<td>1.072 (0.008)</td>
<td>0.000</td>
<td>0.019&lt;sup&gt;g&lt;/sup&gt;</td>
<td>14072.3</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td>PF</td>
<td>0.973</td>
<td>0.973 (0.004)</td>
<td>0.000</td>
<td>0.026&lt;sup&gt;g&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\phi = 1.0
\]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Theo. CMF&lt;sup&gt;a&lt;/sup&gt;</th>
<th>CMF (SD)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Bias</th>
<th>E&lt;sup&gt;c&lt;/sup&gt;</th>
<th>AIC&lt;sup&gt;d&lt;/sup&gt;</th>
<th>MAD&lt;sup&gt;e&lt;/sup&gt;</th>
<th>MSPE&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>LW</td>
<td>0.900</td>
<td>0.903 (0.023)</td>
<td>-0.003</td>
<td>0.354</td>
<td>13736.2</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>CD</td>
<td>1.072</td>
<td>1.072 (0.009)</td>
<td>0.000</td>
<td>0.032&lt;sup&gt;g&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF</td>
<td>0.973</td>
<td>0.972 (0.004)</td>
<td>0.001</td>
<td>0.063</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\phi = 2.0
\]

a, b, c, d, e, f, g - the same notes as those in Table 4; * - LW = Lane Width, CD = Curve Density, PF = Pavement Friction.

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 is not the case in most crash prediction studies. Not all the factors affecting crashes are known or able to be captured by the model. In scenario III, another condition is considered: both curve density and pavement friction are associated with crash risk, but only the lane width is included in the model.

The assumed CMFs for lane width, curve density and pavement friction are 0.90, 1.072 and 0.973, respectively, the same as those in scenario II. The theoretical function of scenario III is the same as that of scenario II, as shown in Equation 16b (now 18 below). In this scenario, the curve density and pavement friction are excluded from the model. So, the model for scenario III is basically same as that of scenario I, as shown in Equation 12 above (now 19 below).

\[
N_{true,i} = 0.0023 \times L_i \times AADT_i \times \exp(-0.105LW_i + 0.070CD_i - 0.027PF_i)
\]

\[
E(\Lambda_i) = \beta_0 \times L_i \times AADT_i^{\beta_1} \times \exp(\beta_2 \times LW_i)
\]

The results for scenario III are shown in Table 6. The derived CMF for lane width in this scenario is also close to the assumed value 0.90. The bias is relatively small and the error is within 1.2 percent in this experiment. Generally, the CMF for lane width in this experiment is reliable.

However, when compared with the results in scenarios I and II, both the bias and error percentage become large in scenario III. That is, when some factors affecting traffic safety are
omitted in the models, the bias for the CMF will be higher. Meanwhile, the MAD and MSPE also increase greatly, indicating the modeling result becomes 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 have been analyzed, and the results are consistent.

Table 6 Results of Scenario III

<table>
<thead>
<tr>
<th>Theo. CMF a</th>
<th>CMF (SD) b</th>
<th>Bias</th>
<th>E c</th>
<th>AIC d</th>
<th>MAD e</th>
<th>MSPE f</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\phi = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.898 (0.014)</td>
<td>0.002</td>
<td>0.211</td>
<td>14395.9</td>
<td>0.75</td>
<td>2.45</td>
</tr>
<tr>
<td>$\phi = 1.0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.890 (0.026)</td>
<td>0.010</td>
<td>1.111</td>
<td>14479.1</td>
<td>0.75</td>
<td>2.49</td>
</tr>
<tr>
<td>$\phi = 2.0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.898 (0.023)</td>
<td>0.002</td>
<td>0.206</td>
<td>13975.8</td>
<td>0.75</td>
<td>2.51</td>
</tr>
</tbody>
</table>

a, b, c, d, e, f - the same notes as those in Table 4.

In scenario III, the assumed CMFs for curve density and pavement friction are close to 1.0. This means the change of one unit of these two variables will have relatively small effect or association on crash risk. In other words, if minor factors are omitted in the SPFs, the result may still be acceptable. However, the bias may become unacceptable if the omitted factors have 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. Then, the CMF for curve density was increased to 1.4, at which point the maximum likelihood algorithm for estimating coefficients failed to converge. Therefore, when major factors are omitted in the SPFs, the CMFs derived may become unreliable. This indicates that the model suffers from the omitted-variable bias (14).

DISCUSSION AND CONCLUSIONS

This research has sought to examine the conditions in which regression models could be used for developing CMFs. CMFs for three variables, lane width, curve density and pavement friction, were assumed and used for generating random crash counts based on a NB modeling structure. Then, CMFs were derived from GLM models using the simulated crash data for three different scenarios. Finally, the modeling results were compared with the assumed true value.

The primary conclusions can be summarized as follows: 1) the CMFs derived from regression models should be unbiased when all factors associated with the crash risk are identical in all segments, except the factors of interest; 2) based on the simulation results, if some factors having minor safety effects are omitted in the SPFs, the derived CMFs are still reliable, but the error and bias become larger; 3) When some factors known to significantly influence crash risk
are omitted, the CMFs derived from regression models become unreliable. According to these findings, safety analysts are recommended to examine whether or not the segments or sites 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), attention should be paid to the use of the CMFs derived from the models.

It is worth mentioning that the CMFs for all the three variables in this paper are assumed to have an exponential relationship with their corresponding variables. This is, in fact, a linear relationship between the predicted crash risk and the changes in some variable (in the logarithmic form). Two discussion points need to be addressed. First, this linear relationship is consistent with the GLM model used in this study. The expected crash mean will always be multiplied by a constant factor when a variable increases by one unit, regardless of the original value of the variable. This might explain the small bias and error percentage observed in this study. Second, this relationship is monotonic. For example, if the CMF for lane width is less than 1.0, the expected crash mean will always decrease when the lane width increases. However, recent research has shown that this may not be the case. Some variables have been shown to have non-monotonic relationships with crash risk (31; 32). Under such conditions, the GLM method may not be applicable, since the CMFs produced from these models will be biased, especially at the boundary conditions. Recently, other statistical models have been proposed to address this problem (22; 33). It is therefore suggested to examine non-monotonic relationships between CMFs and variables using an approach similar to the one documented in this paper. Finally, the factors influencing safety in this study were assumed to be “independent” of each other, which may also not be realistic in practice (2). Further work is needed to examine this issue in the context of regression models.

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 (7; 34). 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.

In conclusion, the work in this paper is a first step on better understanding the development of CMFs using regression models. A solid model and proper assumption of the relationship between variables and their safety effects are the base for developing reliable CMFs when using the regression method. This study started from the simple form of CMFs (i.e., independent of each other, in linear form) and commonly used model (i.e., GLM based on the NB distribution). The sample size, which is assumed to be large enough, will also influence the CMFs. The topic of this paper can be expanded to other areas, such as independence of variables, non-linear relationships between crash risk and variables. Some of the proposed work is ongoing and the results will be documented in future manuscripts.

REFERENCES


