The Poisson Inverse Gaussian (PIG) Generalized Linear Regression Model for Analyzing Motor Vehicle Crash Data

By

Liteng Zha
Graduate Research Assistant
Email: seuzha@gmail.com
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843

Dominique Lord, Ph.D., P.E.
Associate Professor
Telephone: 979-458-3949
Fax: 979-845-6481
E-mail: d-lord@tamu.edu
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843

Yajie Zou*, Ph.D.
Research associate
133B More Hall, Box 352700
Department of Civil and Environmental Engineering
University of Washington, Seattle, WA 98195-2700
Tel: (936) 245-5628 Fax: (206) 543-5965
Email: zouyajie@gmail.com

Submitted for publication

March 31, 2014

* Corresponding author
ABSTRACT

This paper documents the application of the Poisson Inverse Gaussian (PIG) regression model for modeling motor vehicle crash data. The PIG distribution, which mixes the Poisson distribution and Inverse Gaussian distribution, has the potential for modeling highly dispersed count data due to the flexibility of Inverse Gaussian distribution. The objectives of this paper were to evaluate the application of PIG regression model for analyzing motor vehicle crash data and compare the results with Negative binomial (NB) model, especially when varying dispersion parameter is introduced. To accomplish the objectives, both NB and PIG models were developed with fixed and varying dispersion parameters and compared using two datasets. The Texas undivided rural highway segments dataset includes five years of crash data, while the divided highway segments Washington dataset includes four years. The results of this study show that PIG models perform better than the NB models in terms of goodness-of-fit (GOF) statistics. Moreover, PIG models can perform similarly well in capturing the variance of crash to the NB models. Lastly, PIG models demonstrate almost the same prediction performance compared to NB models. Considering the simple form of PIG model and its easiness of applications, PIG model could be used as a potential alternative to the NB model for analyzing crash data.
1. INTRODUCTION

Statistical models are an important component of highway safety research. Over the years, much effort has been devoted to modeling the motor vehicle crash data (Park et al., 2010, Mannering and Bhat, 2014). Because crash data are often characterized by over-dispersion, where the variance is greater than the mean, the Negative Binomial (NB) model (also called Poisson-Gamma) was proposed as a good alternative to the Poisson model for analyzing this kind of datasets. The NB model allows for capturing the extra variation by adding a randomly distributed error term, which is based on the Gamma distribution, as the name implies. Since the introduction of the NB model, it has been widely applied by transportation safety analysts (Miaou, 1994, Milton and Mannering, 1998, Kim et al., 2006, Lord and Bonneson, 2007). For example, the Safety Performance Functions (SPFs) listed in Highway Safety Manual (HSM) were all developed based on NB model (AASHTO, 2010). Moreover, the NB model is the most commonly used in the Empirical Bayesian (EB) method.

More recently within the framework of Generalized Linear models (GLM), many new models have been developed and show potential alternatives to the NB regression model (Park et al., 2010). Each of them has its own characteristics and could provide a better fit compared to the traditional NB model for the corresponding conditions. Such models include, for instance, the Double Poisson (DP) and Conway-Maxwell-Poisson (COM-Poisson) for their flexibility of handling both over-dispersion and under-dispersion (Zou et al., 2012a, Lord et al., 2008b); Random-parameter models (Anastasopoulos and Mannering, 2009; Chen and Tarko, 2014) considering the temporal and spatial correlation for longitudinal data; and the Generalized Additive Model (GAM) which extends the GLM and provides more flexible functional forms (Xie and Zhang, 2008).

The statistical performance of many of the models mentioned above has been shown to be deficient for datasets with large number of zeros and highly skewness. So far, very few studies have examined this issue. A Negative Binomial-Lindley model (NB-L) was introduced by Geedipally et al. (2012) to model crash data characterized by large number of zeros. The NB-L model was demonstrated to provide better statistical performance than the Zero Inflated Negative Binomial model (ZINB) and is more theoretically sound (Lord et al., 2005). However, since the likelihood function for the NB-L does not have a closed form, the parameter estimation based on Markov chain Monte Carlo (MCMC) chain requires intensive computation time.

Zou et al. (2013) proposed the Sichel Generalized Additive Models for Location, Scale and Shape to handle severe dispersion among crash data. The Sichel (SI) distribution is a mixture of Poisson distribution and Generalized Inverse Gaussian distribution (Stein et al., 1987). According to the authors, the SI model provides a better fit than the NB model, especially when the data are highly dispersed. The SI distribution has three parameters which makes its probability density function (PDF) more complicated than the NB model and consequently requiring more computation effort,
As the special type of the SI distribution by setting shape parameter $\gamma = -0.5$, the Poisson Inverse Gaussian (PIG) distribution may also be flexible to handle crash data, but is only characterized by two parameters. In fact, a few studies in areas such as medicine and motor insurance have suggested the PIG as an alternative to the NB model for modeling count data distribution since the PIG distribution has slightly longer tails and larger kurtosis (Willmot, 1987, Shoukri et al., 2004, Dean et al., 1989, Jagger and Elsner, 2012). Besides, the likelihood function can be easily obtained and has nicely closed form, which indicates the estimation of parameters will be quite simple and almost take no time (Stasinopoulos and Rigby, 2012). Despite these nice properties, the application of the PIG model in highway safety (e.g. modeling crash data, hot spot identification) has not be fully investigated.

The objective of this paper is therefore to evaluate the application of the PIG model for analyzing crash data and compare its results with NB model, particularly when a varying dispersion parameter is used. Previous work has shown that models with varying dispersion parameters can have significant effects of common highway safety analyses (Geedipally and Lord, 2008, Lord and Park, 2008, Geedipally and Lord, 2011). To accomplish these objectives, both NB and PIG models were developed and compared using two datasets. The Texas dataset includes five years’ crash data on undivided rural highway segments in Texas, while the Washington dataset includes four years’ crash data on divided highway segments in Washington. The commonly used link function which associated crash with covariates such as daily traffic flow, lane width was chosen. Both models were developed using varying dispersion parameters for the flexibility of accounting for the variation and better statistical fit (Hauer, 2001, Heydecker and Wu, 2001, Miaou and Lord, 2003, Mitra and Washington, 2007, Geedipally and Lord, 2011). Models with fixed dispersion parameters were also listed for comparison purposes. The varying dispersion parameter was modeled as a function of section length, as was suggested by Geedipally and Lord (2011). Statistical fit and prediction performance of the two models were thoroughly studied and compared.

2. BACKGROUND

This section is divided into four parts. Firstly, general derivation of Poisson mixture is briefly discussed. Then, the characteristics of the PIG distribution is described in details. The derivation of empirical Bayesian (EB) framework for the PIG model is also provided, although it is not formally evaluated in this study. It is provided to show how the PIG model can be used efficiently for highway safety applications. Lastly, the log likelihood function for estimating regression parameters is summarized.

2.1. Poisson Mixture

It is commonly accepted that the mean of crash on Segment $i$, $Y_i$, follows a Poisson distribution, which is conditional on its mean $\mu_i$ and independent of all the segments (Miaou and Lord, 2003):
\[ Y_i | \mu_i \sim \text{Poisson}(\mu_i) \quad i = 1, 2, \ldots, n \]  

(1)

With its mean and variance structured in the following regression equation:

\[ \mu_i = E(Y_i | \mu_i) = Var(Y_i | \mu_i) = f(X_i; \beta) = \text{EXP}(X \beta) \]  

(2)

where

- \( f(.) \) = Link functional form (log in this case);
- \( X \) = the vector of covariates;
- \( \beta \) = regression parameters associated with corresponding covariates.

For crash data, the equal-dispersion is too restrictive. Very often, the conditional variance will exceeds the conditional mean (over-dispersion), which is likely to result from positive contagion and unobserved heterogeneity. An error term \( \varepsilon_i \) is introduced to \( \mu_i \) and

\[ \text{EXP}(X^T \beta + \varepsilon_i) = \mu_i \text{EXP}(\varepsilon_i) = \mu_i \nu_i \]  

(3)

The precise form of the mixed Poisson distribution depends on the specific distribution that \( \nu_i \) follows. Let \( g(\nu_i) \) be the probability density function (PDF) of \( \nu_i \) then the marginal distribution for \( Y_i \) can be obtained by integrating out \( \nu_i \):

\[ P(Y_i = y_i | \mu_i) = \int f(y_i | \mu_i, \nu_i) g(\nu_i) d\nu_i \]  

(4)

If \( \nu_i \) is assumed to be Gamma distributed with mean equal to 1 and variance \( 1/\phi \), the Poisson-Gamma (or Negative Binomial) regression model is then obtained.

### 2.2. Poisson Inverse-Gaussian Regression Model

For the PIG distribution, \( \nu_i \) in Equation (4) is assumed to be independent of all covariates and follows an Inverse Gaussian distribution with mean equal to 1 and shape parameter \( 1/\tau (\nu_i \sim IG(1, 1/\tau)) \). Then, the PDF for \( \nu_i \) can be written as (Stasinopoulos and Rigby, 2007):

\[ g(\nu_i) = (2\pi\nu_i^{3})^{-0.5} e^{-(\nu_i^{-1})^{2}/2\tau} \nu_i, \quad \nu_i > 0. \]  

(5)

where \( \tau = \text{Var}(\nu_i); \ E(\nu_i) = 1 \).
Eventually, the PIG distribution, denoted by PIG(µᵢ, τᵢ), is given by:

\[
P(yᵢ | µᵢ, τᵢ) = \left( \frac{2\alpha_i}{\pi} \right)^{\frac{1}{2}} \frac{µ_i^{\alpha_i} e^{\frac{1}{2} τ_i} K_{\alpha_i}(y_i)}{(α,τ_i)^{\frac{1}{2}} y_i!}
\]  \[ (6) \]

where \( \alpha_i = \frac{1}{τ_i^2} + \frac{2µ_i}{τ_i} \);

\[
K_{α}(t) = \frac{1}{2} \int_0^\infty x^{α-1} e^{-\frac{1}{2}(x+t^{-1})} dx, \text{ which is the modified Bessel function of the third kind (Stasinopoulos and Rigby, 2007).}
\]

The mean for the PIG distribution is given as:

\[
E(Yᵢ) = E\{E(Yᵢ | µᵢνᵢ)\} = E(µᵢνᵢ) = µᵢ
\]  \[ (7) \]

The variance for PIG distribution is given as:

\[
Var(Yᵢ) = Var\{E(Yᵢ | µᵢνᵢ)\} + E\{Var(Yᵢ | µᵢνᵢ)\} = µᵢ + τᵢ \bar{µ}^2
\]  \[ (8) \]

### 2.3. Empirical Bayesian Method Based on the PIG Model

The EB method has been widely applied to identify hotspots by ranking crash-prone locations. Previously, a SI-based EB method has been developed by Zou et al. (2013). Since the PIG model is a special case of the SI model, similar to the dispersion parameter in the NB model, the scale parameter \( τ \) of the PIG model can be used to obtain reliable EB estimates. A PIG modeling framework to obtain EB estimates is described below.

According to Hauer (1997), let \( K \) be the observed number of crashes which is Poisson distributed, and \( κ \) be the expected crash count, the EB estimator of \( κ \) is:

\[
\hat{κ} = wE(κ) + (1-w)K
\]  \[ (9) \]

where \( \hat{κ} \) denotes the EB estimate of the expected number of crashes. \( E(κ) \) can be estimated by the PIG model. The weight factor \( w \) is given as (Hauer, 1997):

\[
w = \frac{1}{1 + \frac{VAR(κ)}{E(κ)}}
\]  \[ (10) \]

The weight factor \( w \) is a function of mean and variance of \( κ \) and is always a number between 0 and 1. If we let \( κ \) be inverse Gaussian distributed, then the resulting \( K \)
follows a PIG distribution. For the PIG model, the weight factor can be estimated as a function of \( E(\kappa) \) and scale parameter \( \tau \), which is

\[
w = \frac{1}{1 + \tau E(\kappa)}
\]  

(11)

As shown above, it is very convenient to apply the PIG model to calculate the EB estimates. Since the scope of this paper is to propose the PIG model as an alternative model for analyzing over-dispersed crash data, the performance of the PIG-based EB method will be examined in a future study.

2.4. Maximum Likelihood Estimating Methods

One advantage of the PIG distribution is that its likelihood function has a closed form. Therefore, regression parameters can be easily obtained through the Maximum Likelihood Estimating (MLE) Method. The probability generating function for PIG \((\mu, \tau)\) is given as (Dean et al., 1989):

\[
P(z) = \sum_{y=0}^{\infty} p(y)z^y = \exp(\tau^{-1}[1 - \{1 - 2\mu(z - 1)^{1/2}\}])
\]  

(12)

Then probabilities can be calculated recursively,

\[
p(0) = \exp(\tau^{-1}[1 - (1 + 2\mu)^{1/2}]),
\]

\[
p(1) = \mu(1 + 2\mu)^{-1/2} p(0),
\]

\[
p(y) = \frac{2\tau\mu}{1 + 2\tau\mu} (1 - \frac{3}{2y})p(y - 1) + \frac{\mu^2}{1 + 2\tau\mu} \frac{1}{y(y - 1)} p(y - 2), y = 2, 3, ...
\]  

(13)

The log-likelihood function for a random sample of observations \((y_i, x_i)\) can thus be derived as:

\[
l(\beta, \tau) = \sum_{i=1}^{n} \left\{ \log\left(\frac{1}{y_i!}\right) + \log p(y_i) + I(y_i > 0) \sum_{j=0}^{y_i-1} \log[(y_i+1)\frac{p_j(y_i+1)}{p_j(y)}] \right\}
\]  

(14)

More information for the derivatives for the log likelihood function can be found at Dean et al. (1989).
3. METHODOLOGY

This section described regression functions that were used for parameter estimation of the proposed models and procedures for examining the goodness-of-fit (GOF) as well as the prediction performance.

The functional form for regression models is listed as follows:

\[ \mu_i = L_i \times \beta_0 \times F_i^{\beta_1} \times e^{\sum_{j} \beta_j x_{ij}} \]  

(15)

where,

\( \mu_i \) = mean number of crashes per year for segment \( i \);

\( F_i \) = average annual daily traffic on segment \( i \);

\( L_i \) = length of segment \( i \);

\( X_{ij} \) = explanatory variables (e.g., lane width, segment length, etc.) for the segment \( i \); and,

\( \beta_0, \beta_1, \beta_2... \) = estimated coefficients.

Basically, it links the crash mean and the explanatory variables and is the most commonly used functional form. Previous studies have shown that the dispersion parameter can also be modeled as a function of the explanatory variables to explain more variation. The link functional form for dispersion parameters was also investigated by authors (Hauer, 2001, Heydecker and Wu, 2001, Miaou and Lord, 2003, Mitra and Washington, 2007). Considering that Equation (15) contains all the covariates (Mitra and Washington, 2007), we simply modeled the dispersion parameters of NB and PIG models as the function of segment length in the format recommended by Geedipally and Lord (2011):

\[ \alpha; \tau = \gamma_0 L_i^{\gamma_1} \]  

(16)

where,

\( \alpha; \tau \) = dispersion parameter for NB model and PIG model;

\( L_i \) = length of segment \( i \); and,

\( \gamma_0, \gamma_1 \) = estimated coefficients.

The nonlinear relationship provides more flexibility to capture the variance of the data. Our preliminary analyses also showed that it gave better GOF than other forms (e.g. linear: \( \gamma_0 L_i \), inverse: \( \gamma_0 / L_i \)).

For comparing the statistical fit of the NB and PIG models, both fixed and varying dispersion parameters were examined using each entire dataset. For models with varying
dispersion parameters, the distribution of the dispersion parameters as well as the crash-mean relationship from the models was investigated to determine the ability of the PIG and NB models in accounting for the variation of the data.

The evaluation of the predicted performance was accomplished in two stages. Firstly, four samples including 80% of the entire data were randomly selected from each dataset for model estimation. Then, the remaining 20% was used to predict crash data based on the obtained regression model (Lord et al., 2008b).

4. DATA DESCRIPTION

This section described two datasets that were collected as part of the NCHRP research project leading by Lord et al. (2008a) for the purpose of this study. The Texas dataset includes five years’ crash data on undivided rural highway segments in Texas, while the Washington dataset includes four years’ crash data on divided highway segments in Washington State.

As mentioned above, distributions of crash data are often right-skewed with lots of zeros. We define the skewness as:

$$Skewness = \frac{\sum_{i=1}^{N} (Y_i - \mu)^3}{(N-1)s^3}$$

where $Y_i$ is the recorded crashes over the study period, $\mu$ is the mean of the crashes, $s$ is the standard deviation and $N$ is the number of road segments.

Negative values and positive values for skewness indicate the data is left-skewed and right-skewed, respectively. As a rule of thumb, the distribution is considered highly skewed when the absolute value of skewness is larger than one (Bulmer, 1979).

4.1. Texas Data

The Texas crash dataset included 4,253 crashes that occurred on 1,499 undivided rural highway segments in Texas over a five-year period from 1997 to 2001. It has been explored by former studies (Lord et al., 2008b, Zou et al., 2012b). Figure 1 shows the distribution of the crash data. It has a long tail with skewness 6.2. Among the 1,499 segments, 37% of them reported zero crashes. The variance over mean ratio is 11.4, which indicates high over-dispersion.

Note that the segments used in this database have not been modified during the study period and are no less than 0.1 mile in length. Other variables included lane width, shoulder width, curve density as well as segment length. The summary statistics for the explanatory variables are displayed in Table 1 below.
Figure 1 Crash Distribution for Texas Data.

Table 1 Summary Statistics of the Explanatory Variables of Texas Data.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT (Veh/day)</td>
<td>42.00</td>
<td>24800.00</td>
<td>6613.61</td>
<td>4010.01</td>
</tr>
<tr>
<td>Lane Width(ft)</td>
<td>9.75</td>
<td>16.50</td>
<td>12.57</td>
<td>1.59</td>
</tr>
<tr>
<td>Shoulder Width(ft)</td>
<td>0</td>
<td>40.00</td>
<td>9.96</td>
<td>8.02</td>
</tr>
<tr>
<td>Curve Density(/Mile)</td>
<td>0</td>
<td>18.07</td>
<td>1.43</td>
<td>2.35</td>
</tr>
<tr>
<td>Segment Length(ft)</td>
<td>0.10</td>
<td>6.27</td>
<td>0.55</td>
<td>0.67</td>
</tr>
<tr>
<td>Median Width (ft)</td>
<td>1</td>
<td>240</td>
<td>47.71</td>
<td>28.87</td>
</tr>
</tbody>
</table>

* S.D. denotes standard deviation

4.2. Washington Data

The Washington state dataset included 2,282 crashes that happened on 476 divided highway segments during the four-year study period. Also, the segments used in this data base have not been modified during the study period and are no less than 0.1 mile in length. Figure 2 shows the distribution of crashes. Similar to the Texas data, the distribution is positively skewed with the skewness 3.4. A variance over mean ratio of 6.0 indicates over-dispersion of the crash data. Explanatory variables are the same as those of Texas data and are summarized in Table 2.
Table 2 Summary Statistics of the Explanatory Variables of Washington Data.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT (Veh/day)</td>
<td>3187</td>
<td>61947</td>
<td>15625.86</td>
<td>10271.42</td>
</tr>
<tr>
<td>Lane Width (ft)</td>
<td>11</td>
<td>17</td>
<td>12.15</td>
<td>0.81</td>
</tr>
<tr>
<td>Shoulder Width(ft)</td>
<td>0</td>
<td>40</td>
<td>9.58</td>
<td>2.45</td>
</tr>
<tr>
<td>Curve Density(/Mile)</td>
<td>0</td>
<td>20</td>
<td>2.84</td>
<td>3.45</td>
</tr>
<tr>
<td>Segment Length(ft)</td>
<td>0.1</td>
<td>2.76</td>
<td>0.41</td>
<td>0.42</td>
</tr>
<tr>
<td>Median Width (ft)</td>
<td>4</td>
<td>620</td>
<td>67.06</td>
<td>73.45</td>
</tr>
</tbody>
</table>

5. MODELING RESULTS

This section described the comparisons of the results for statistical fit as well as prediction performance of NB and PIG models. Both models were estimated using the GAMLSS Library of R Package (Stasinopoulos and Rigby, 2007).

5.1. Comparison of Goodness-of-fit (GOF)

Table 3 and Table 4 summarize the estimation results of NB and PIG models using Texas Data and Washington Data, respectively. Variables in the models were chosen using the stepwise method and only those significantly at the 0.05 significance level were kept in the regression model.
### Table 3 Modeling Results for the Texas Data.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>NB</th>
<th>PIG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fixed dispersion parameter</td>
<td>Varying dispersion parameter</td>
</tr>
<tr>
<td>$\ln \beta_0$</td>
<td>-7.949(0.406)*</td>
<td>-7.6(0.402)</td>
</tr>
<tr>
<td>$\ln(AADT)$ ((\beta_1))</td>
<td>0.975(0.044)</td>
<td>0.947(0.043)</td>
</tr>
<tr>
<td>Lane Width ((\beta_2))</td>
<td>-0.053(0.017)</td>
<td>-0.062(0.017)</td>
</tr>
<tr>
<td>Shoulder Width ((\beta_3))</td>
<td>-0.01(0.003)</td>
<td>-0.011(0.003)</td>
</tr>
<tr>
<td>Curve Density ((\beta_4))</td>
<td>0.067(0.012)</td>
<td>0.078(0.013)</td>
</tr>
<tr>
<td>$\ln \alpha ; \ln \tau$</td>
<td>-0.94(0.092)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>--</td>
<td>-1.144(0.096)</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>--</td>
<td>-0.49(0.091)</td>
</tr>
<tr>
<td>Global Deviance</td>
<td>5122.772</td>
<td>5096.408</td>
</tr>
<tr>
<td>AIC:</td>
<td>5134.772</td>
<td>5110.408</td>
</tr>
<tr>
<td>BIC:</td>
<td>5166.647</td>
<td>5147.596</td>
</tr>
</tbody>
</table>

* Numbers in the Parenthesis is the Standard Error. Same for the Remaining of the Paper.

### Table 4 Modeling Results for the Washington Data.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>NB</th>
<th>PIG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fixed dispersion parameter</td>
<td>Varying dispersion parameter</td>
</tr>
<tr>
<td>$\ln \beta_0$</td>
<td>-3.03(0.531)</td>
<td>-2.973(0.531)</td>
</tr>
<tr>
<td>$\ln(AADT)$ ((\beta_1))</td>
<td>0.453(0.054)</td>
<td>0.445(0.054)</td>
</tr>
<tr>
<td>Shoulder Width ((\beta_2))</td>
<td>-0.038(0.017)</td>
<td>-0.038(0.017)</td>
</tr>
<tr>
<td>Curve Density ((\beta_3))</td>
<td>0.068(0.011)</td>
<td>0.078(0.013)</td>
</tr>
<tr>
<td>$\ln \alpha ; \ln \tau$</td>
<td>-1.248(0.132)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>--</td>
<td>-1.546(0.183)</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>--</td>
<td>-0.372(0.153)</td>
</tr>
<tr>
<td>Global Deviance</td>
<td>2180.556</td>
<td>2174.921</td>
</tr>
<tr>
<td>AIC:</td>
<td>2190.556</td>
<td>2186.921</td>
</tr>
<tr>
<td>BIC:</td>
<td>2211.383</td>
<td>2211.913</td>
</tr>
</tbody>
</table>

Two conclusions can be made from Tables 3 and 4. Firstly, NB and PIG models provided similar estimates. For the Texas data, for instance, both models showed that AADT and curve density positively associated with the crash frequency while lane width and shoulder width negatively associated with crash frequency. The estimates are reasonable since the increase of AADT increases the exposure for crashes and higher curve density indicates more risk. Comparatively, larger lane width and shoulder width were associated with lower crash risk.

More specifically, 1% increase in AADT for instance, can result in 0.959% and 0.471% (according to the varying dispersion parameter PIG model) increase of crash frequency for the Texas data and Washington data, respectively. One unit increase in curve density leads to 7.90% and 7.57% (\(e^{0.076} - 1\), \(e^{0.073} - 1\), according to the varying dispersion
parameter PIG model) increase of the crash frequency in Texas data and Washington data, respectively. Similar analysis can be done for other explanatory variables.

Secondly, the PIG model showed better GOF than the NB model, as expected, and models with varying dispersion parameters exhibited a slightly better fit than the fixed-dispersion model, as noted by Miaou and Lord (2003). Generally, the global deviance, Akaike information criterion (AIC) and Bayesian information criterion (BIC) values consistently indicated that the PIG models with varying dispersion parameter provided better statistical performance. This may be explained by the better flexibility of Inverse Gaussian distribution in fitting the crash data than the Gamma distribution. The varying of dispersion parameter generally captured more variability of the crash data. One exception exists in the Washington data, PIG model with fixed dispersion parameter has the smallest BIC value.

5.2. Distributions of Varying Dispersion Parameters
The distributions of the varying dispersion parameters for NB and PIG models are exhibited in Figure 3, with the mean highlighted by dash lines. As seen, the dispersion parameters for the Texas data had wider variation than those for the Washington data. For the Texas data, the dispersion parameters for NB and PIG models occurred most frequently in the interval $[0.4, 0.5]$ with an average equal to 0.50 and 0.58, respectively.

![Figure 3 Distribution of Dispersion Parameter of NB and PIG Model for Texas Data ((a), (b)) and Washington Data ((c), (d)).](image)
Comparatively, those parameters in the fixed dispersion parameter model were 0.39 and 0.45. For the Washington data, dispersion parameter for the NB model occurred within the interval [0.3, 0.4] with the average 0.35 while that for the PIG model occurred most frequently in the range [0.4, 0.5] with the average 0.38. Similarly to the Texas data, the dispersion parameters in the fixed dispersion parameter models have smaller values, which were 0.29 and 0.31 for the NB and PIG models, respectively.

5.3. Crash Variance-Mean Relationship
The relationship between crash variance and crash mean estimated from the above models is given in Figure 4. It clearly showed that for both two datasets, the variance-mean curves of NB and PIG models were perfectly overlapped with almost the same tendency. The Spearman’s rank correlation coefficients between the estimated variance were found to be 0.999 and 0.998, respectively. This indicates that the NB model and PIG with the varying dispersion parameter have similarly performance in accounting for the variation of the crash data for each segment during the observation periods.

![Figure 4 Crash Variance Versus Crash Mean for (a) Texas Data and (b) Washington Data.](image)
5.4. Comparison of Prediction Performance

Analysis of prediction performance was accomplished using two-step approach. Firstly, four samples including 80% of the entire data were randomly selected from each dataset for model estimation. Based on the obtained regression model, the remaining 20% was used to predict crash data. It should be noted that for the remaining of the paper, all the models were developed using varying dispersion parameters, which generally were demonstrated to provide better statistical fit.

Comparison of prediction accuracy for NB and PIG model was based on Mean absolute deviance (MAD) and mean squared predictive error (MSPE) (Cameron and Trivedi, 1998, Lord et al., 2008b). Smaller values of these two measurements indicate better prediction performance. Detailed results are listed through Table 5 to Table 8.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Sample1</th>
<th>Sample2</th>
<th>Sample3</th>
<th>Sample4</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln \beta_0$</td>
<td>-7.956(0.442)</td>
<td>-7.955(0.45)</td>
<td>-7.44(0.467)</td>
<td>-7.354(0.446)</td>
<td>--</td>
</tr>
<tr>
<td>ln(AADT) ($\beta_1$)</td>
<td>0.983(0.047)</td>
<td>0.983(0.048)</td>
<td>0.942(0.051)</td>
<td>0.919(0.048)</td>
<td>--</td>
</tr>
<tr>
<td>Lane Width ($\beta_2$)</td>
<td>-0.06(0.019)</td>
<td>-0.059(0.019)</td>
<td>-0.073(0.02)</td>
<td>-0.062(0.018)</td>
<td>--</td>
</tr>
<tr>
<td>Shoulder Width ($\beta_3$)</td>
<td>-0.012(0.004)</td>
<td>-0.01(0.004)</td>
<td>-0.009(0.004)</td>
<td>-0.009(0.004)</td>
<td>--</td>
</tr>
<tr>
<td>Curve Density ($\beta_4$)</td>
<td>0.085(0.014)</td>
<td>0.073(0.014)</td>
<td>0.071(0.015)</td>
<td>0.07(0.015)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>-1.223(0.113)</td>
<td>-1.223(0.108)</td>
<td>-1.113(0.11)</td>
<td>-1.191(0.11)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.57(0.105)</td>
<td>-0.482(0.104)</td>
<td>-0.523(0.104)</td>
<td>-0.46(0.104)</td>
<td>--</td>
</tr>
<tr>
<td>AIC:</td>
<td>4083.733</td>
<td>4049.029</td>
<td>4022.877</td>
<td>4054.920</td>
<td>--</td>
</tr>
<tr>
<td>BIC:</td>
<td>4119.358</td>
<td>4084.654</td>
<td>4058.502</td>
<td>4090.544</td>
<td>--</td>
</tr>
<tr>
<td>MAD</td>
<td>1.744</td>
<td>1.797</td>
<td>1.907</td>
<td>1.951</td>
<td>1.850</td>
</tr>
<tr>
<td>MSPE</td>
<td>10.438</td>
<td>8.531</td>
<td>23.034</td>
<td>22.300</td>
<td>16.076</td>
</tr>
</tbody>
</table>
### Table 6 Modeling Results of The PIG Model Using Texas Data.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Sample1 Estimate</th>
<th>Sample2 Estimate</th>
<th>Sample3 Estimate</th>
<th>Sample4 Estimate</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln \beta_0$</td>
<td>-8.008(0.455)</td>
<td>-8.024(0.466)</td>
<td>-7.535(0.484)</td>
<td>-7.493(0.461)</td>
<td>--</td>
</tr>
<tr>
<td>$\ln(\text{AADT})$</td>
<td>0.989(0.049)</td>
<td>0.994(0.05)</td>
<td>0.954(0.052)</td>
<td>0.933(0.05)</td>
<td>--</td>
</tr>
<tr>
<td>$\text{Lane Width (} \beta_2 \text{)}$</td>
<td>-0.059(0.019)</td>
<td>-0.061(0.019)</td>
<td>-0.073(0.02)</td>
<td>-0.061(0.019)</td>
<td>--</td>
</tr>
<tr>
<td>$\text{Shoulder Width (} \beta_3 \text{)}$</td>
<td>-0.012(0.004)</td>
<td>-0.009(0.004)</td>
<td>-0.009(0.004)</td>
<td>-0.009(0.004)</td>
<td>--</td>
</tr>
<tr>
<td>$\text{Curve Density (} \beta_4 \text{)}$</td>
<td>0.084(0.014)</td>
<td>0.071(0.014)</td>
<td>0.07(0.015)</td>
<td>0.067(0.015)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>-1.128(0.127)</td>
<td>-1.123(0.12)</td>
<td>-1.013(0.124)</td>
<td>-1.095(0.122)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.613(0.119)</td>
<td>-0.532(0.115)</td>
<td>-0.57(0.116)</td>
<td>-0.494(0.114)</td>
<td>--</td>
</tr>
<tr>
<td>AIC:</td>
<td>4073.816</td>
<td>4033.106</td>
<td>4014.796</td>
<td>4044.105</td>
<td>--</td>
</tr>
<tr>
<td>BIC:</td>
<td>4109.441</td>
<td>4068.731</td>
<td>4050.421</td>
<td>4079.730</td>
<td>--</td>
</tr>
<tr>
<td>MAD</td>
<td>1.745</td>
<td>1.797</td>
<td>1.906</td>
<td>1.951</td>
<td>1.850</td>
</tr>
<tr>
<td>MSPE</td>
<td>10.459</td>
<td>8.538</td>
<td>22.884</td>
<td>22.376</td>
<td>16.064</td>
</tr>
</tbody>
</table>

### Table 7 Modeling Results of the NB Model Using Washington Data.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Sample1 Estimate</th>
<th>Sample2 Estimate</th>
<th>Sample3 Estimate</th>
<th>Sample4 Estimate</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln \beta_0$</td>
<td>-3.209(0.593)</td>
<td>-2.922(0.596)</td>
<td>-3.064(0.586)</td>
<td>-2.698(0.607)</td>
<td>--</td>
</tr>
<tr>
<td>$\ln(\text{AADT})$</td>
<td>0.459(0.06)</td>
<td>0.428(0.06)</td>
<td>0.461(0.06)</td>
<td>0.425(0.061)</td>
<td>--</td>
</tr>
<tr>
<td>$\text{Shoulder Width (} \beta_2 \text{)}$</td>
<td>-0.03(0.02)</td>
<td>-0.023(0.018)</td>
<td>-0.039(0.019)</td>
<td>-0.047(0.019)</td>
<td>--</td>
</tr>
<tr>
<td>$\text{Curve Density (} \beta_4 \text{)}$</td>
<td>0.082(0.014)</td>
<td>0.071(0.015)</td>
<td>0.061(0.014)</td>
<td>0.084(0.015)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>-1.679(0.236)</td>
<td>-1.443(0.2)</td>
<td>-1.396(0.197)</td>
<td>-1.602(0.212)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.448(0.186)</td>
<td>-0.328(0.173)</td>
<td>-0.068(0.191)</td>
<td>-0.433(0.172)</td>
<td>--</td>
</tr>
<tr>
<td>AIC:</td>
<td>1705.584</td>
<td>1746.767</td>
<td>1739.855</td>
<td>1755.452</td>
<td>--</td>
</tr>
<tr>
<td>BIC:</td>
<td>1729.209</td>
<td>1770.408</td>
<td>1763.496</td>
<td>1779.093</td>
<td>--</td>
</tr>
<tr>
<td>MAD</td>
<td>2.987</td>
<td>2.413</td>
<td>2.583</td>
<td>2.431</td>
<td>2.604</td>
</tr>
<tr>
<td>MSPE</td>
<td>34.460</td>
<td>10.760</td>
<td>12.372</td>
<td>14.737</td>
<td>18.082</td>
</tr>
</tbody>
</table>
Table 8 Modeling Results of The PIG Model Using Washington Data.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Sample1 Estimate</th>
<th>Sample2 Estimate</th>
<th>Sample3 Estimate</th>
<th>Sample4 Estimate</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln $\beta_0$</td>
<td>-3.402(0.62)</td>
<td>-3.186(0.626)</td>
<td>-3.349(0.611)</td>
<td>-2.95(0.638)</td>
<td>--</td>
</tr>
<tr>
<td>ln(AADT) ($\beta_1$)</td>
<td>0.474(0.063)</td>
<td>0.452(0.064)</td>
<td>0.489(0.062)</td>
<td>0.447(0.064)</td>
<td>--</td>
</tr>
<tr>
<td>Shoulder Width ($\beta_2$)</td>
<td>-0.025(0.02)</td>
<td>-0.018(0.018)</td>
<td>-0.036(0.019)</td>
<td>-0.041(0.019)</td>
<td>--</td>
</tr>
<tr>
<td>Curve Density ($\beta_3$)</td>
<td>0.079(0.014)</td>
<td>0.068(0.014)</td>
<td>0.057(0.013)</td>
<td>0.079(0.014)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>-1.614(0.259)</td>
<td>-1.371(0.219)</td>
<td>0.218(0.124)</td>
<td>-1.523(0.231)</td>
<td>--</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.47(0.214)</td>
<td>-0.353(0.196)</td>
<td>0.206(0.116)</td>
<td>-0.445(0.195)</td>
<td>--</td>
</tr>
<tr>
<td>AIC:</td>
<td>1704.240</td>
<td>1744.174</td>
<td>1734.915</td>
<td>1753.146</td>
<td>--</td>
</tr>
<tr>
<td>MAD</td>
<td>2.987</td>
<td>2.410</td>
<td>2.583</td>
<td>2.402</td>
<td>2.596</td>
</tr>
</tbody>
</table>

It was found that the average MAD and MPSE values (highlighted in bold) for NB and PIG models were almost the same, which demonstrated that PIG model can do as well as NB model in terms of prediction performance. It is also interesting to notice that for each sample the PIG models unexceptionally have the smaller AIC and BIC values than NB models, although the difference is not very large.

6. DISCUSSION

This study examined the application of a special case of the SI distribution: the PIG distribution in the analyzing crash data. The PIG model provided slightly better statistical fit than the NB model (the statistical fit for both models can be further improved by varying the dispersion parameters) and almost the same prediction performance as the NB model. This is probably attributed to the more flexibility of the Inverse Gaussian distribution than the Gamma distribution. Thus, PIG model was demonstrated to be a potential alternative of NB model in analyzing crash data.

One obvious advantage of PIG model is that its likelihood function can be easily obtained and has nicely closed form. MLE is then applicable for estimating the parameters. In other words, just as NB model, little computation effort is needed to apply the PIG models. Comparatively, recent introduced models such as NB-L model which relies on MCMC chain for parameter estimation may require a few hours simulation before converging (Lord et al., 2008b, Geedipally et al., 2012). Actually, at the end of the paper written by Zou et al. (2012b), the potential of applying PIG rather than SI was also addressed mainly for its simplicity.

Although PIG model is more flexible than NB model, its performance for crash data with small sample mean (e.g. large number of zeros) was not fully investigated. By closely examining the two datasets, it was not difficult to find that the proportions of zero crash
were not very high. Only 37% and 15% of the segments in the Texas dataset and Washington data reported zero crash. A simulation study should be done to further explore the performance of the PIG model in handling crash data with lots of zeros.

7. SUMMARY

This paper documented the application of PIG model for analyzing motor vehicle crash data. As a special case for SI distribution by setting shape parameter $\gamma = -0.5$, the PIG distribution has the potential to handle highly dispersed data and is easy to compute. The objectives of this paper were to evaluate the application of PIG regression model for analyzing motor crash data and compare its results with the NB model, particularly when a varying dispersion parameter is used. To accomplish the objectives, both NB and PIG models were developed and compared using two datasets with fixed and varying dispersion parameters. Commonly used link functional forms for crash mean as well as varying dispersion parameters were chosen. Statistical fit of NB and PIG models as well as the prediction performance was thoroughly studied and compared. Note that a PIG modeling framework to obtain EB estimates was also derived, although it was not evaluated due to the limited scope of this study.

The results of this study showed that in terms of GOF statistics, the PIG model was slightly better than NB model with both varying and fixed dispersion parameters. Moreover, the model can perform similarly well in capturing the variance of crash and demonstrates almost the same prediction performance compared to NB model. Considering its easiness of computation and convenience of adoption in the EB analysis framework, PIG model turns out to be a potential alternative to NB model for analyzing motor crash data.
REFERENCES


