GOODNESS-OF-FIT TESTING FOR ACCIDENT MODELS WITH LOW MEANS

Zhirui Ye
Research Scientist, Western Transportation Institute, Montana State University,
Bozeman, MT, USA, e-mail: jared.ye@coe.montana.edu

Yunlong Zhang
Associate Professor, Zachry Department of Civil Engineering, Texas A&M University,
College Station, TX, USA, e-mail: yzhang@civil.tamu.edu

Dominique Lord
Associate Professor, Zachry Department of Civil Engineering, Texas A&M University
College Station, TX, USA, e-mail: d-lord@tamu.edu

Submitted to the 3rd International Conference on Road Safety and Simulation,
September 14-16, 2011, Indianapolis, USA

ABSTRACT

The modeling of relationships between motor vehicle crashes and underlying factors has been investigated for more than three decades. Recently, many highway safety studies have documented the use of Poisson regression models, negative binomial (NB) regression models or both. Pearson's X^2 and the scaled deviance (G^2) are two common test statistics that have been proposed as measures of goodness-of-fit (GOF) for Poisson or NB models. Unfortunately, transportation safety analysts often deal with crash data that are characterized by low sample mean values. Under such conditions, the traditional test statistics may not perform very well.

This study has two objectives. The first objective is to examine the accuracy and reliability of traditional test statistics for the GOF of accident models subjected to low sample means. The second objective intends to identify a superior test statistic for evaluating the GOF of accident prediction models. For Poisson models, this paper proposes a better yet easy to use test statistic (Power-Divergence) that can be applied for almost all sample mean values, except when the mean value is extremely low, for which no traditional test statistic can be accurate. For Poisson-Gamma models, this study demonstrates that traditional test statistics are not accurate and robust. A more complex method (grouped G^2) proposed in a previous study is recommended. Guidance on the use of the grouped G^2 methods is further provided. Examples using observed data are used to help illustrate the performance of different test statistics and support the findings of this study.

Keywords: crash data, generalized linear model, goodness-of-fit, power-divergence.

INTRODUCTION

The modeling of relationships between motor vehicle crashes and underlying factors, such as traffic volume and highway geometric features has been investigated for more than three decades. The statistical models (sometimes referred to as crash prediction models) from which these relationships are developed can be used for various purposes, including predicting crashes on transportation facilities and determining which variables significantly influence crashes. Recently, many highway safety studies have documented the use of Poisson regression models (Joshua and Garber, 1990; Miaou et al., 1992; Ivan and Bernardo, 2000; Lord and Bonneson, 2007), negative binomial (NB) regression models (Miaou and Lum, 1993; Poch and Mannering, 1996; Miaou and Lord, 2003; Maycock and Hall, 1984; Lord et al., 2005) or both (Miaou, 1994; Maher and Summersgill, 1996). With the Poisson or Poisson-Gamma (or NB) models, the relationships between motor vehicle crashes and explanatory variables can then be developed by means of the Generalized Linear Model (GLM) framework.

Pearson's X^2 and the scaled deviance (G^2) are two common test statistics that have been proposed as measures of GOF for Poisson or NB models (Maher and Summersgill, 1996). Statistical software (e.g., SAS) also uses these two statistics for assessing the GOF of a GLM (SAS Institute Inc., 1999). Unfortunately, transportation safety analysts often deal with crash data that are subjected to low sample mean values. Under such conditions, the traditional test statistics may not perform very well. This has been referred to in the highway safety literature as the low mean problem (LMP). The study by Sukhatme (1938) concluded that, "for samples from a Poisson distribution with mean as low as one, Pearson's X^2 test for goodness of fit is not good." In the field of traffic safety, this issue was first raised by Maycock and Hall (1984) and further discussed by Maher and Summersgill (1996), Fridstrom et al., (1995), and Agrawal and Lord (2006). Wood (2002) proposed a more complex technique, the grouped G^2 method, to solve this problem. The grouped G^2 method is based on the knowledge that through grouping, the data become approximately normally distributed and the test statistics follow a χ^2 distribution. Some issues regarding this method are discussed in the third section. It should be noted that the comparison of different models can be achieved by means of Akaike's Information Criterion (AIC) (Akaike, 1974) or Bayesian Information Criterion (BIC) (Schwarz, 1978). However, similar to the previous studies (Maher and Summersgill, 1996; Wood, 2002; Agrawal and Lord, 2006), this research intends to study statistics for the GOF of a given model (either Poisson model or NB model); thus, we mainly focused on the study of the statistics of X^2, G^2 and the proposed statistic (Power-Divergence).

This study expands on the work of Wood (2002) and has two objectives. The first objective is to examine the accuracy and reliability of traditional test statistics for the GOF of GLMs subjected to low sample means. The second objective intends to identify a superior test statistic for evaluating the GOF of crash prediction models. The study is accomplished by first theoretically deriving the problems related with these traditional tests. Observed data are then used to demonstrate the problems noted in the first part of the paper.

This paper is divided into five sections. The second section describes the characteristics of Poisson and NB models used in traffic crash modeling. The third section provides an analysis and comparison of different GOF test statistics for the Poisson and NB models. Observed crash data are used for this analysis. In the fourth section, several important issues related to the GOF test statistics are discussed. The last section summarizes the key findings of this study.

STATISTICAL MODELS

GLMs represent a class of fixed-effect regression models for dependent variables (McCullagh and Nelder, 1989), such as crash counts in traffic accident models. Common GLMs include linear regression, logistic regression, and Poisson regression. Given the characteristics of motor vehicle collisions (i.e., random, discrete, and non-negative independent events), stochastic modeling methods need to be used over deterministic methods. The two most common stochastic modeling methods utilized for analyzing motor vehicle crashes are the Poisson and the NB regression models. For these models, the relationship between traffic accidents and explanatory variables is established through a loglinear function (i.e., canonical link or linear predictor). For example, to establish the crash-flow relationship at intersections, the fitted model can follow the form $\mu = \beta_0 \times F_1^{\beta_1} \times F_2^{\beta_2}$, where μ is the estimated number of crashes, F_1 and F_2 are the entering AADTs (Average Annual Daily Traffic) for major and minor approaches, and $\beta_0, \beta_1, \beta_2$ are the estimated coefficients. This fitted model can thus be used for predicting crashes for different flow values.

Poisson Regression Model

The Poisson regression model aims at modeling a crash count variable Y, which follows a Poisson distribution with a parameter (or mean) μ. The probability that the number of crashes takes the value y_i on the *i*th entity is $P(Y_i = y_i) = f_{Y_i}(y_i; \mu_i) = \frac{\mu_i^{y_i} \times e^{-u_i}}{y_i!}, \quad \mathrm{i} = 1,2,...,\mathrm{n}$. For a Poisson distribution, the variance is equal to the mean.

The systematic portion of the model involves the explanatory variables $x_1, x_2, ..., x_m$, such as traffic volumes, highway geometrics, v/c (volume/capacity) ratios and so on. The model is then established through a linear predictor η. This predictor is usually a linear function of the logarithm of the explanatory variables in traffic crash models: $\eta = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$, where β_i is the Poisson regression coefficient for the *i*th explanatory variable x_i. The coefficients are estimated based on observed data. Finally, the model is estimated through a logarithm link function $\eta_j = g(\mu_j) = \log(\mu_j)$ (Myers et al., 2002).

Negative Binomial Regression Models

Although Poisson regression models are rather simple, crash data often exhibit overdispersion, meaning that the variance is greater than the mean. The NB regression models are thus used for modeling such data. The NB regression models have the same forms of linear predictor and logarithm link function as the Poisson regression models, except that the response variable Y follows a NB distribution, in which the probability mass function (pmf) is defined as follows:

$$P(Y_i = y_i) = f_{Y_i}(y_i;\phi;\mu_i) = \frac{\Gamma(\phi + y_i)}{\Gamma(\phi)\times\Gamma(y_i + 1)}\left(\frac{\phi}{\mu_i + \phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i},$$

where $\Gamma(\cdot)$ is a Gamma function, and ϕ is the inverse dispersion parameter. The relationship between the variance and the mean of NB distribution is presented as $Var(Y_i) = \mu_i + \mu_i^2/\phi$. The inverse dispersion parameter is usually assumed to be fixed and can be estimated from observed data using the method of moments or the (Bootstrapped) maximum likelihood (Anscombe, 1949; Fisher, 1941; Zhang et al., 2007). However, recent research have shown that the inverse dispersion parameters may be related to the explanatory variables (Miaou and Lord, 2003; Mitra and Washington, 2007).

GOODNESS-OF-FIT TEST STATISTICS

GOF tests use the properties of a hypothesized distribution to assess whether or not observed data are generated from a given distribution (Read and Cressie, 1988). The most well-known GOF test statistics are Pearson's X^2 and the scaled deviance (G^2). Pearson's X^2 is generally calculated as follows: $X^2 = \sum_{i=1}^{n}\left[\frac{y_i - \mu_i}{\sigma_i}\right]^2$, where y_i is the observed data, μ_i is the true mean from the model, and σ_i is the error and is usually represented by the standard deviation of y_i. The scaled deviance is calculated as twice the difference between the log-likelihood under the maximum model and the log-likelihood under the reduced (or unsaturated) model: $G^2 = 2(\log L_{max.} - \log L_{red.})$ (Wood, 2002).

Previous research has shown that both the Pearson's X^2 and G^2 statistics are not χ^2 distributed under low sample mean conditions (Maycock and Hall, 1984; Maher and Summersgill, 1996; Wood, 2002; Fridstrom et al., 1995; Agrawal and Lord, 2006). To solve this problem, Maher and Summersgill (1996) proposed a test statistic ($G^2/E(G^2)$) for GOF tests. Wood (2002) showed that this test still failed with low sample mean values. Wood (2002) then suggested a grouped G^2 test statistic for solving this problem. The development of the grouped G^2 is based on the knowledge that by increasing the mean value, the data are approximately normally distributed and the statistics follow a χ^2 distribution. This method first determines an appropriate group size r, which is the minimum grouping size. The raw data are then grouped so that each observation is in a group of size at least as large as r. Additional details about the other steps can be found in Wood (2002).

There are some issues with this method, however, that need to be addressed with the method proposed by Wood (2002). First, the grouping size may vary from group to group with a minimum grouping size, which is determined by the sample mean of a Poisson model or the critical mean values in a NB model, as defined in Wood (2002). Thus, it is possible that changing grouping sizes while maintaining the same minimum grouping size may lead to different testing results. Second, through grouping, the sample size will be smaller, and that may become an issue especially when the grouping size is not small. Thus, as commented by Wood, a compromise has to be made between strong grouping (which ensures that the Chi-square assumption for the distribution of the test statistic holds) and weak grouping (which allows to test against a richer alternative hypothesis). Finally, the grouped G^2, which includes five steps, is not a simple procedure for practitioners or average transportation safety analysts who frequently analyze crash data.

To summarize, several GOF test statistics have been proposed to evaluate the fit of models, but their performance and complexity vary greatly. Therefore, simple but accurate and reliable alternative test statistics are highly desirable to account for the LMP commonly observed in crash studies.

In Wood's study (2002), a simple criterion to assess whether or not a test statistic is appropriate for testing the GOF of regression models is to examine the test statistic's performance for a single distribution (Poisson or NB) with known parameters. For this criterion, the grouped G^2 method was developed to improve the normality of observations and allow the mean and variance of the G^2 statistic (for low mean μ values) to be close to 1 and 2 (χ_1^2 distributed), respectively. Similarly in this study, we examine the mean and variance of different statistics under a single distribution context to judge their appropriateness for the GOF of GLM.

Test Statistics for Poisson Models

<u>Characteristics of Statistical Tests</u>

The most common test statistics are Pearson's X^2 and the scaled deviance (G^2). For a Poisson model, the variance is equal to the mean and Pearson's X^2 is presented below:

$$X^2(\mu;n) = \sum_{i=1}^{n}[\frac{y_i - \mu_i}{\sigma_i}]^2 = \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{\mu_i} \tag{1}$$

The scaled deviance for a Poisson model is (Maher and Summersgill, 1996)

$$G^2(\mu;n) = \sum_{i=1}^{n} 2[y_i \log(\frac{y_i}{\mu_i}) - (y_i - \mu_i)] \tag{2}$$

In this paper, we investigate other test statistics for the GOF test of the Poisson model, especially when it is characterized by low sample mean values. This research draws from some other work in the statistical literature.

Cressie and Read (1984 & 1988) incorporated the Pearson's X^2 and G^2 statistics into a family of "Power-Divergence Statistics" (PD_λ, $\lambda \in R$). In this family, each member PD_λ is the sum of deviance between the observed and expected counts:

$$PD_\lambda = \sum_i a_\lambda(y_i, \mu_i) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{n} [y_i((\frac{y_i}{\mu_i})^\lambda - 1)], \quad -\infty < \lambda < +\infty \tag{3}$$

where a_λ denotes the distance function. Different values of λ lead to different GOF statistics (Cressie and Read, 1984 & 1988; Baggerly, 1998), such as the Pearson's X^2 statistic when $\lambda = 1$, the Freeman-Tukey statistic $F^2 = PD_{\lambda=-1/2} = 4\sum_{i=1}^{n}(\sqrt{y_i} - \sqrt{\mu_i})^2$ when $\lambda = -1/2$ (Freeman and Tukey, 1950), and the Neyman-modified X^2 statistic $NM^2 = PD_{\lambda=-2} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{y_i}$ when $\lambda = -2$ (Neyman, 1949). The Power-Divergence statistic can be also written as (Cressie and Read, 1989)

$$PD_\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{n} [y_i((\frac{y_i}{\mu_i})^\lambda - 1)] - \lambda(y_i - \mu_i)], \quad -\infty < \lambda < +\infty \tag{4}$$

Hence, when $\lambda \to 0$, the power divergence leads to the G^2 statistic (Cressie and Read, 1989).

Cressie and Read (1988) recommended $\lambda = 2/3$, with which the statistic $PD_{\lambda=2/3}$ will be approximated by the χ^2 distribution in many situations and give the most reasonable power for GOFs. When $\lambda = 2/3$, the test statistic of Power-Divergence becomes $PD_{\lambda=2/3} = \sum_{i=1}^{n} [\frac{9y_i}{5}((\frac{y_i}{\mu_i})^{2/3} - 1) - \frac{6}{5}(y_i - u_i)]$, as derived from Equation 4.

GOF tests using different statistics rest on the assumption that the statistics follow an approximate χ^2 distribution that has a mean of 1 and a variance of 2. Thus, to evaluate a test statistic for GOF tests, we can investigate how well its components follow a χ^2 distribution. With this criterion, different test statistics can be compared and evaluated.

Pearson's X^2, the G^2, Power-Divergence with $\lambda = 2/3$ ($PD_{\lambda=2/3}$), and the Freeman-Tukey statistic $F^2 = 4\sum_{i=1}^{n}(\sqrt{y_i} - \sqrt{\mu_i})^2$ (Freeman and Tukey, 1950) are used for the examination of the fit of χ^2 distributions. In the case that crash data have zero counts at some locations, the Neyman-modified X^2 goes to infinity and is therefore excluded from the comparison analysis. Figure 1 shows the mean and variance of the components of those four statistics, for the Poisson mean μ less than 10. The following equations show the calculations of mean and variance of the Pearson's X^2 statistic, given a known Poisson mean value (μ):

$$E(X^2)=\sum_{y=0}^{\infty} X^2 f_Y(k;\mu) \tag{5}$$

$$V(X^2)=\sum_{y=0}^{\infty} [X^2 - E(X^2)]^2 f_Y(k;\mu) \tag{6}$$

where $f_Y(k;\mu)$ is the pmf of Poisson distributions and k is the number of occurrence of an event. The mean and variance of other statistics over different μ values can be calculated in this way.

The comparisons are first conducted for μ values varying from 1 to 10. They are shown in Figure 1. From this figure, Pearson's X^2 has a mean value ($E(X^2)$) of 1 for all μ values, but its variance ($V(X^2)$) is greater than 2. With the decrease of μ, the variance increases. Thus, for low μ conditions, Pearson's X^2 is not reliable and as a result, tends to overestimate GOF values. In fact, $V(X^2)$ is equal to $2+1/\mu$ and this has also been described in the study by Wood (2002). The mean of the scaled deviance ($E(G^2)$) is slightly larger than 1 (when μ >1) and moves toward 1 as μ rises; the variance ($V(G^2)$) increases from less than 1 to around 2.4 and then decreases toward 2. The Freeman-Tukey statistic does not have a good fit of χ^2 distributions even when $\mu=10$. The mean and the variance of the $PD_{\lambda=2/3}$ statistic, however, are rather close to 1 and 2 respectively. The components of the $PD_{\lambda=2/3}$ statistic fit χ^2 distributions almost perfectly as long as u >1. Therefore, the $PD_{\lambda=2/3}$ is recommended for GOF tests for $\mu \in [1,10]$.

Figure 2 shows the comparison of mean and variance of X^2, G^2, and $PD_{\lambda=2/3}$, for $0.1 \le \mu < 1$. It can be observed that $E(G^2)$ varies from 0.47 to 1.15, while $E(PD)$ increases from 0.7 to 0.98. Overall, $E(PD)$ is more stable based on the rate of increase and is much closer to 1.0 than $E(G^2)$. For $\mu \ge 0.3$, the difference between $E(PD)$ and $E(X^2)$, which is exactly 1, is very small and negligible. $V(G^2)$ is always less than 2 and even less than 1 given $\mu < 0.7$; $V(PD)$ has the same tendency as $V(X^2)$, but is more stable and gets close to 2.0 even when μ is as small as 0.3, while $V(X^2)$ stays above 3.0 at $\mu = 1$. It can be also seen that $V(PD)$ performs like a compromise between $V(X^2)$ and $V(G^2)$. From the above comparisons, for $\mu \in [0.3, 1]$, the components of $PD_{\lambda=2/3}$ are approximately χ^2 distributed and $PD_{\lambda=2/3}$ performs better than the other statistics. For $\mu < 0.3$, no statistic is reliable for GOF tests, and practitioners may consider turning to the more complicated grouped G^2 method.

Based on Figures 1 and 2, $PD_{\lambda=2/3}$ is better than the other statistics and its components generally fit χ^2 distributions well for $\mu > 0.3$. Pearson's X^2 is slightly better than G^2 for $\mu > 3$, but even when $\mu = 10$, Pearson's X^2 and the G^2 are not satisfactory, with means and variances of *($E(X^2)=1.00$, $V(X^2)=2.10$)* and *($E(G^2)=1.02$, $V(G^2)= 2.09$)*, respectively.

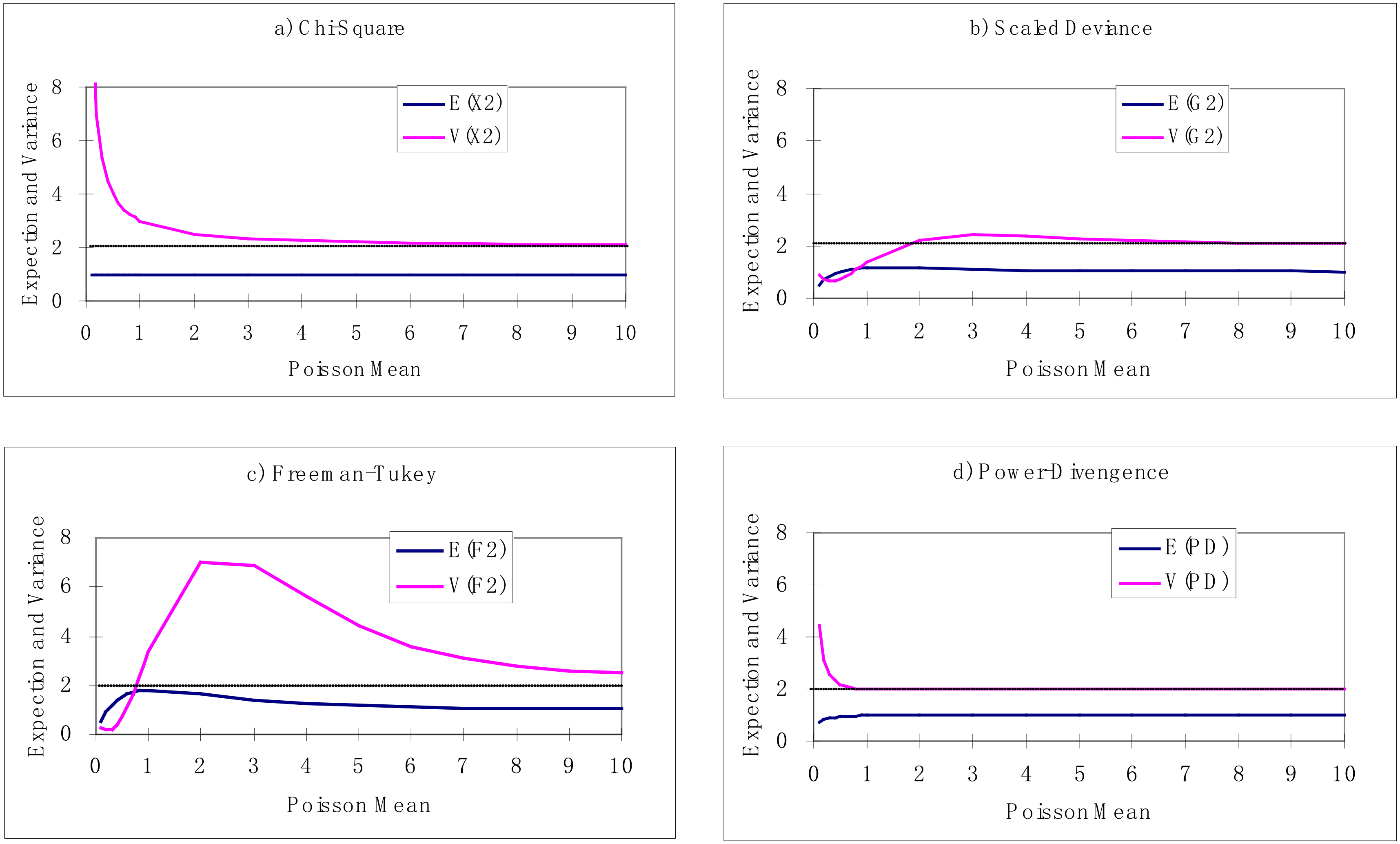

Figure 1 Mean and variance of the components of different test statistics for $0 < u < 10$

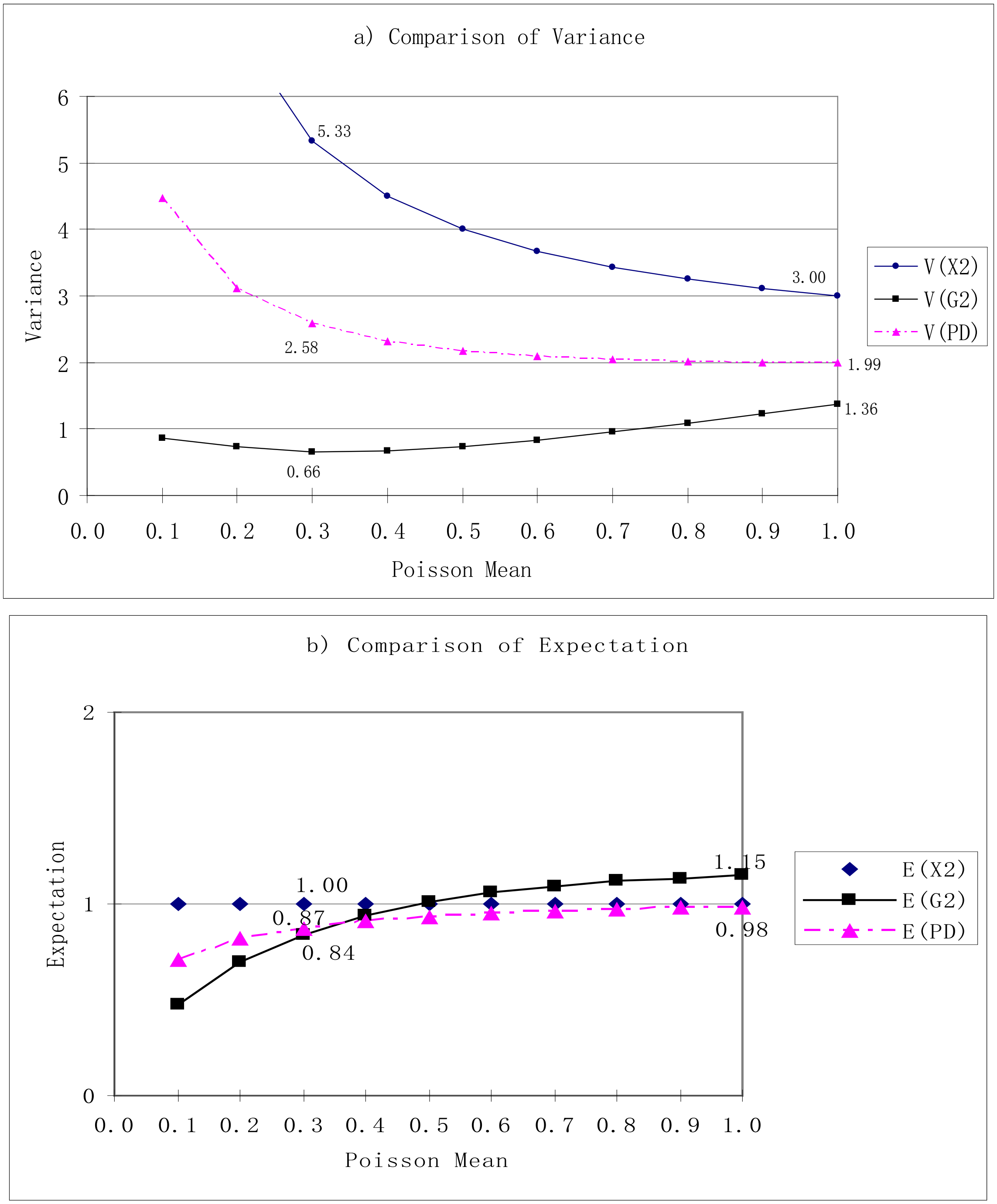

Figure 2 Mean and variance of components of different test statistics for $0.1 \le u \le 1$

Example Applications with Observed Data

To show how different GOF test statistics affect the fit of Poisson models, two examples using observed crash data are provided. It is worth noting that the core of the study is to statistically investigate the performance of different test statistics for GOF under low mean conditions. The following data are used as examples to help support the findings from statistical investigation, but not to serve as alternative approach to investigate their performance.

For the first example, the data were collected at 59 four-legged unsignalized intersections in 1991 in Toronto, Ontario (Lord, 2000). The dataset includes the number of crashes and entering AADT for the major and minor approaches at each site. Both Poisson and NB GLM were used for modeling this dataset, but the NB model converged to Poisson, with an inverse dispersion parameter that tended towards infinity (Lord and Bonneson, 2007). The mean of this dataset is 0.97. The variance is roughly the same as the mean. Thus, a Poisson GLM could be used for modeling this dataset. The functional form $\mu = \beta_0 \times F_1^{\beta_1} \times F_2^{\beta_2}$ is used for the prediction of the number of crashes. As stated by Lord (2006), it is the most common functional form used by transportation safety analysts for modeling crash data at intersections. The outputs of the fitted model are shown in Table 1. It can be seen that all coefficients are still significant even at the significance level of 0.01.

Table 1 Modeling outputs of the Poisson model

Coefficients	Est. Value	Std. Error	*z* value	*Pr(>|z|)*
β_0	2.3439E-06	4.2895	-3.022	0.0025
β_1	0.8175	0.3145	2.599	0.0093
β_2	0.6348	0.2349	2.7303	0.0069

Pearson's X^2, G^2, $PD_{\lambda=2/3}$, and F^2 are used for the GOF test of this Poisson model. The results of the GOF tests are summarized in Table 2. The $PD_{\lambda=2/3}$ statistic has a lower GOF value and correspondingly a higher *p*-value than the Pearson's X^2 statistic. The GOF value of G^2 is higher than Pearson's X^2. The F^2 statistic has the lowest *p*-value. To explain their differences, Table 2 also lists the mean and variance of those test statistics given the Poisson mean $\mu = 0.97$. The mean and variance of the distribution of the test statistics can also be seen from Figure 1 or Figure 2. It is clear that the components of the $PD_{\lambda=2/3}$ statistic are rather close to a χ^2 distribution. For the Pearson's X^2 statistic, $E(X^2) = 1$ and $V(X^2) = 3.03$. The variance $V(X^2)$ is larger than 2 and may have overestimated the GOF value given $E(X^2) = 1$. For the G^2 statistic, although $V(G^2) = 1.23 < 2$, the mean $E(G^2) = 1.14$ is higher than 1 and can also result in overestimations of GOF values. Similarly, the F^2 statistic will also overestimate GOF values.

Table 2 Results of GOF tests for the Poisson model

Statistics	X^2	PD	G^2	F^2
GOF value	52.71	51.76	59.85	93.41
Degrees of Freedom	56	56	56	56
p-value	0.60	0.64	0.34	0.00
Expectation*	1.00	0.98	1.14	1.81
Variance**	3.03	1.99	1.23	3.20

*: The means of test statistics when the Poisson mean is 0.97.
**: The variances of test statistics when the Poisson mean is 0.97.

For the second example, the data were collected at 88 frontage road segments in the State of Texas (Lord and Bonneson, 2007). The dataset includes the number of serious injury crashes (KAB or K=Fatal, Injury Type A – incapacitated, and Injury Type B – non-incapacitated), segment length, and AADT. The mean of this dataset is 1.386 and the variance is 1.642. Both Poisson and NB GLM were used for modeling this dataset, but the NB model converged to Poisson, with an inverse dispersion parameter that tended towards infinity (Lord and Bonneson, 2007). The functional form $\mu = \beta_0 * L * F^{\beta_1}$ was used for the prediction of the number of crashes, where L represents the segment length and F is the AADT. The modeling results are shown in Table 3. It can be seen that both coefficients are significant at the significance level of 0.01.

Table 3 Modeling outputs of the Poisson model

Coefficients	Est. Value	Std. Error	z value	$Pr(>\|z\|)$
β_0	0.01536	0.8374	-4.987	6.14e-07
β_1	0.5874	0.1195	4.916	8.82e-07

Again, Pearson's X^2, G^2, $PD_{\lambda=2/3}$, and F^2 are used for the GOF test of this Poisson model. As can be seen from Table 4, the GOF testing results are consistent with those of the first example, which does not warrant further discussion.

Table 4 Results of GOF tests for the Poisson model

Statistics	X^2	PD	G^2	F^2
GOF value	104.84	103.01	116.08	168.87
Degrees of Freedom	86	86	86	86
p-value	0.08	0.10	0.02	0.00
Expectation*	1.00	0.99	1.14	1.75
Variance**	2.81	1.99	1.70	4.76

*: The means of test statistics when the Poisson mean is 1.386.
**: The variances of test statistics when the Poisson mean is 1.386.

Test Statistics for Negative Binomial Models

Characteristics of Statistical Tests

For NB distributions, the variance can be calculated as $Var(Y_i) = \mu_i + \mu_i^2 / \phi$. Thus, the Pearson's X^2 statistic becomes $X^2(\mu;n) = \sum_{i=1}^{n}[\frac{y_i - \mu_i}{\sigma_i}]^2 = \sum_{i=1}^{n}\frac{[y_i - \mu_i]^2}{\mu_i + \mu_i^2 / \phi}$. Based on the definition of the scaled deviance (Wood, 2002), the G^2 statistic for a NB model is calculated by $G^2(\mu;\phi;n) = 2\sum_{i=1}^{n}[\phi \log(\frac{\mu_i + \phi}{y_i + \phi}) + y_i \log(\frac{y_i(\mu_i + \phi)}{\mu_i(y_i + \phi)})]$.

To show the accuracy and reliability of the Pearson's X^2 and G^2 statistics for GOF tests, the components of these statistics are examined again, using the same kind of calculations shown in Equations 5 and 6, in which $f_Y(\cdot)$ is now the pmf of the NB distribution. Note that the Power-Divergence statistics were not used as test statistics in this study for the NB distribution, since they do not exist in the statistical literature. The mean and variance of the Pearson's X^2 and G^2 statistics with different parameter settings are shown in Figure 3. The NB mean (μ) varies from 0 to 10; the inverse dispersion parameters (ϕ) are 1, 3 and 5, respectively. It can be observed that ϕ has a great effect on the distributions of those two statistics. For the Pearson's X^2 statistic, the smaller the inverse dispersion parameter, the larger the *$V(X^2)$* value, given a known NB mean value. The components of the Pearson's X^2 statistic do not fit χ^2 distributions, as *$V(X^2)$* is generally much larger than 2 for low μ values. *$V(X^2)$* is still larger than 3 even when $\mu = 10$ and $\phi = 5$. Therefore, the Pearson's X^2 statistic will underestimate the degree of fit (*p*-value) and tend to reject fitted models more easily in practice. For the G^2 statistic, *$V(G^2)$* may increase or decrease drastically for $\mu < 1$, then gradually stabilizes depending on ϕ. When μ is as high as 10 and $\phi = 1$, the variance *$V(G^2)$* is still not quite stable.

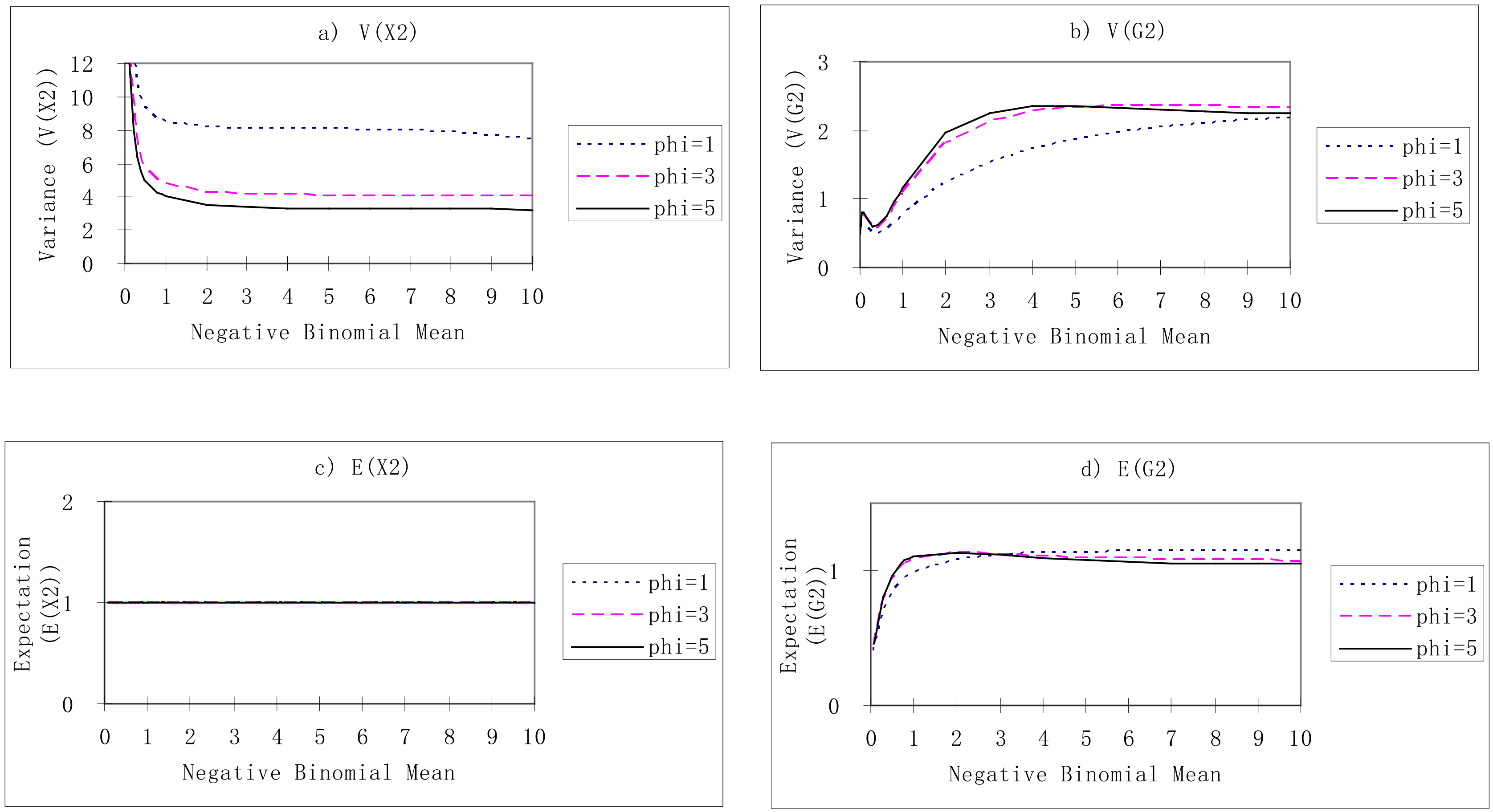

Figure 3 Mean and variance of components of the X^2 and G^2 statistics

With the increase of ϕ, the μ value for $V(G^2)$ to become stable decreases. For example, when $\phi = 3$, $V(G^2)$ becomes stable when μ is around 4, and when $\phi = 5$, $V(G^2)$ will be relatively stable when μ is around 3. $E(G^2)$ is generally greater than 1 for $\mu > 1$ and less than 1 for $0 < \mu < 1$. [Important note: the inverse dispersion parameter is assumed to be properly estimated. As discussed by Lord (2006), the inverse dispersion parameter can become misestimated as the sample mean values decrease and the sample size becomes small.]

Overall, both Pearson's X^2 and G^2 statistics are not quite accurate and reliable for the GOF test of NB models with low sample means, especially when the crash data are highly overdispersed (ϕ is small). As a result, the authors recommend the use of the grouped G^2 method for the GOF test of NB models. An example is given below to show the differences between GOF test statistics for NB models.

Example Applications with Observed Data

An annual crash-flow dataset was collected from 255 signalized 3-legged intersections in Toronto, Ontario (Lord, 2000). This dataset includes the number of serious injury crashes and entering AADTs for the major and minor approaches at each intersection. The crash counts are overdispersed with a mean of 1.43 and a variance of 3.49. A NB regression model was thus used for the modeling of this dataset. The functional form $\mu = \beta_0 \times F_1^{\beta_1} \times F_2^{\beta_2}$ was again used for the prediction of the number of crashes. The results of the fitted model are summarized in Table 5. All the coefficients are significant at the significant level of 0.01. The inverse dispersion parameter was estimated to be 2.76.

Table 5 Modeling outputs of the Negative Binomial model

Coefficients	Est. Value	Std. Error	*z* value	*Pr(>|z|)*
β_0	7.988E-07	2.0122	-6.978	3.00E-12
β_1	1.0241	0.1951	5.249	1.53E-07
β_2	0.4868	0.0821	5.926	3.10E-09

Pearson's X^2, G^2 and the grouped G^2 were used for evaluating the GOF test of the NB model. According to the grouping rules in (Wood, 2002), the minimum grouping size for the dataset is equal to 2, and the expression for calculating the grouped G^2 is $Grouped\_G^2 = 2\sum_{i=1}^{n} r_i[\phi \log(\frac{\mu_i + \phi}{y_i + \phi}) + y_i \log(\frac{y_i(\mu_i + \phi)}{\mu_i(y_i + \phi)})]$, where r_i is the grouping size for the ith group.

The results of GOF tests are summarized in Table 6. The degrees of freedom are 252 for Pearson's X^2 and G^2, and 125 for the grouped G^2. All three test statistics accepted the fitted model at the significance level of 0.05. The grouped G^2 statistic and the Pearson's X^2 statistic have the highest and lowest *p*-values, respectively. The table also shows the expectations and variances of the components of Pearson's X^2 and G^2 statistics, given the known parameters

(μ and ϕ). It can be seen that $E(X^2)$ is 1 and $V(X^2)$ is 4.63. $V(X^2)$ is much larger than 2 and this has caused the overestimation of GOF values. Thus, the *p*-value (0.09) of the X^2 is lower than the actual value. $E(G^2)$ and $V(G^2)$ with low NB mean (μ) values are shown in Figure 4 for $f = 2.76$. When μ is around 1.43, $E(G^2)$ is higher than 1, which may have resulted in the overestimation of GOF values and underestimation of the power of fit; $V(G2)$ is very unstable for low μ values. The *p*-value of the grouped G^2 statistic is slightly higher than that of G^2. This is expected since the G^2 statistic has underestimated the true *p*-value. Thus, this example shows that the grouped G^2, although more complicated than the traditional methods, provides better results for the GOF test of NB models.

Table 6 Results of GOF tests for the Negative Binomial model

Statistics	X^2	G^2	$Grouped\_G^2$
GOF value	282.20	269.80	136.46
Degrees of Freedom	252	252	125
p-value	0.09	0.21	0.23
Expectation*	1	1.12	N/A
Variance**	4.63	1.42	N/A

*: The means of test statistics when the NB mean is 1.43 and the inverse dispersion parameter is 2.756.
**: The variances of test statistics when the NB mean is 1.43 and the inverse dispersion parameter is 2.756.

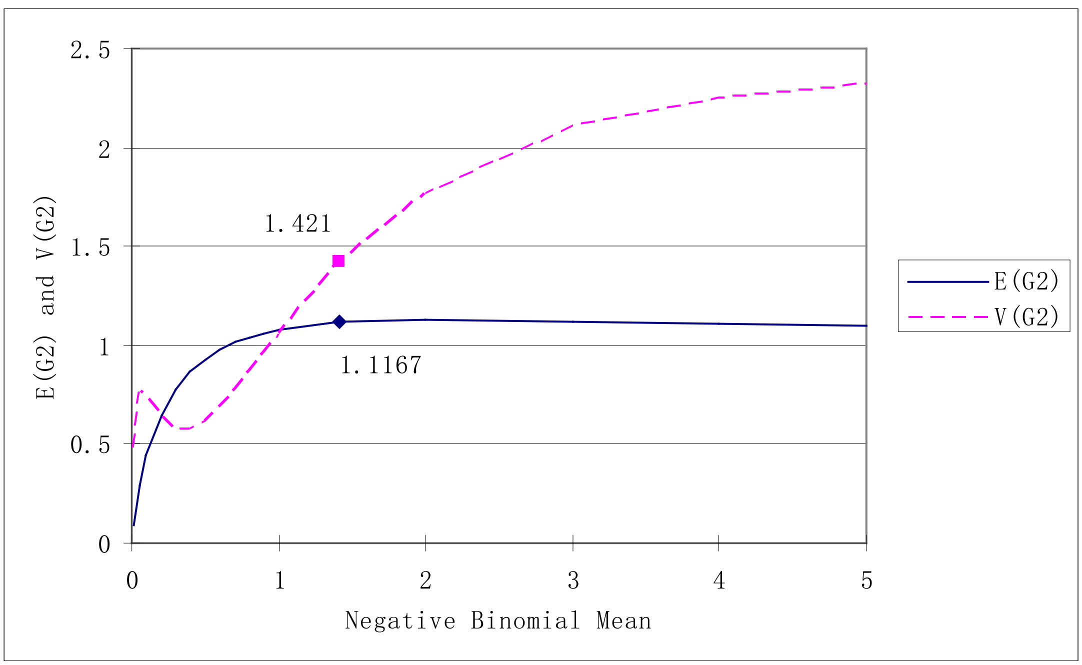

Figure 4 $E(G^2)$ and $V(G^2)$ versus NB mean with $\phi = 2.756$

DISCUSSION

The results of this study show that the Pearson's X^2 statistic tends to overestimate GOF values for low μ values, since *V(X²)* are larger than 2. This is because the components (i.e., $\frac{(y_i - \mu_i)^2}{\mu_i}$ for Poisson models) will be inflated when the predicted values (μ_i) are low. For instance, with the observed crash dataset in the first case, the Poisson model predicted 1.02 crashes per year for one of the intersections. However, 4 crashes were observed at that intersection. The contribution to X^2 would be $(4-1.02)^2/1.02 = 8.71$ and larger than the nominal value. The phenomenon explains why $V(X^2)>2$ for low μ values.

Undoubtedly, for Poisson regression models, the Power-Divergence statistic ($PD_{\lambda=2/3}$) follows an approximate χ^2 distribution and is the best test statistic for measuring the GOF for these models. This statistic performs better than the other three statistics for almost all μ values, except when μ is very low. However, when μ is very small, no test statistics can provide accurate and stable results of GOF tests. This statistic is preferred to the Pearson's X^2 statistic for all cases. For $\mu < 1$, the variance of $PD_{\lambda=2/3}$ statistic (*V(PD)*) performs like a compromise between $V(X^2)$ and $V(G^2)$, and contributes to more accurate and stable GOF tests.

From Figures 1 and 3, it is also observed that the performance of Pearson's X^2 and G^2 becomes worse with the increase in overdispersion. The Poisson model is a special case of the NB model, in which the inverse dispersion parameter is infinite. Therefore, the estimation of the inverse dispersion parameter from observed data will affect the results of GOF tests. It should be noted that the traditional estimators of the inverse dispersion parameter do not have accurate and stable estimations under low mean conditions, as described above (Lord, 2006). For NB models, both Pearson's X^2 and G^2 do not have accurate results of GOF tests, especially under low sample mean conditions. Under such conditions, the grouped G^2 method is recommended, as it will provide better results for GOF tests of NB models.

The results of this study provide guidance on the use of the grouped G^2 method. Based on the curves of G^2 illustrated in Figure 1, it is found that the G^2 method or the grouped G^2 method is an appropriate test statistic only when the grouped mean is 1.5 or higher. Theoretically, the grouped G^2 method can be used for samples with extreme low means (e.g. less than 0.3). However, when grouping a sample with a low mean value to achieve a grouped mean of 1.5 or higher, the grouped sample size will be significantly reduced, which may lead to issues associated with small samples. For NB regression models, the problem becomes more complex as the minimum grouped mean is determined by the inverse dispersion parameter (ϕ). The recommended minimum means (or group means) for different inverse dispersion parameters are shown in Figure 5. The minimum mean decreases when ϕ increases. For ϕ less than 1, the minimum mean increases sharply with a decreasing ϕ. Thus, when using the grouped G^2 method, the grouped mean is suggested to meet the requirements presented in this figure. With the increase of the inverse dispersion parameter towards infinite (Poisson model), the recommended minimum mean decreases slowly to approximately 1.5.

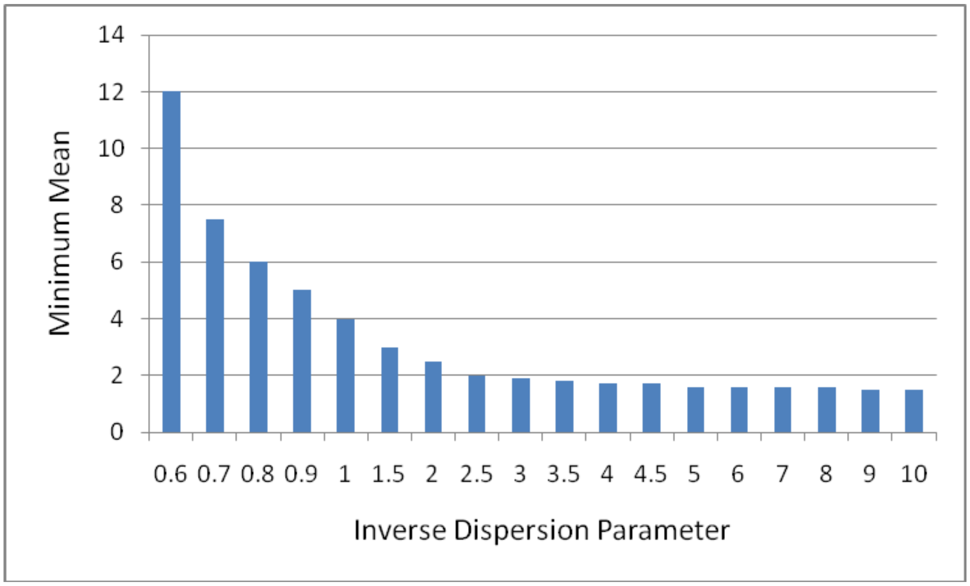

Figure 5 Recommended Minimum Means versus Inverse Dispersion Parameter of the NB model

CONCLUSIONS AND FUTURE WORK

The Poisson and NB regression models are the two most commonly used types of models for analyzing traffic crashes. These models help establish the relationship between traffic crashes (response variable) and traffic flow, highway geometrics, and other explanatory variables. To evaluate their statistical performance, GOF tests need to be used. Since crash data are often characterized by low sample mean values and it has been found that traditional GOF statistics do not perform very well under these conditions. Consequently, there was need to determine whether alternative GOF statistics could be used for data characterized by low sample mean values.

The objectives of this paper were to examine the performance of test statistics for evaluating the GOF of crash models, propose better statistics, and provide useful recommendations for GOF tests for data characterized by low sample mean values. For Poisson models, this paper introduced a test statistic ($PD_{\lambda=2/3}$) that is superior to the traditional statistics. The study showed that the $PD_{\lambda=2/3}$ statistic has accurate GOF tests even when the sample mean is as low as 0.3. For NB distributions with low sample mean values, this paper found that the traditional statistics do not have accurate estimates of the power of fit. Under such conditions, the more complex grouping method is recommended as a remedy. For better illustrations, three examples using observed crash data were used to show the differences of test statistics in GOF tests for Poisson and NB models. Further work should be done to investigate and improve the GOF test of NB models, since this type of model is more often used for modeling crash data.

In the statistical literature, some researchers (Gurtler and Henze, 2000; Spinelli and Stephens, 1997; Baringhaus and Henze, 1992; Kim and Park, 1992; Kocherlakota and Kocherlakota, 1986;

Baglivo et al., 1992) have also examined the performance of GOF test statistics for the Poisson model. Examples include the Cramer-von Mises test statistic and the Kolmogorov-Simirnov test statistic, both well-known for testing the GOF for continuous distributions (Henze and Klar, 1995). Future work can be also conducted to examine whether these test statistics for Poisson models developed using data characterized by low sample mean values perform well. Moreover, Carota (2007) extended the power-divergence to a Bayesian nonparametric context, under which the power-divergence may be an appropriate test statistic for testing NB regression models. Future work can also be conducted in this regard. However, the complexity of the power-divergence with a Bayesian extension can be a barrier for transportation analysts even if it performs well for low mean conditions.

ACKNOWLEDGEMENTS

The authors greatly acknowledge the comments provided by Dr. Graham Wood from the Department of Statistics at Macquarie University in Sydney, Australia.

REFERENCES

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6) 716–723.

Anscombe, F.J. (1949). The statistical analysis of insect counts based on the negative binomial distributions. *Biometrics* 5, 165-173.

Baggerly, K.A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika*, 85(3) 535-547.

Baglivo J., Olivier, D., Pagano, M. (1992). Methods for exact goodness-of-fit tests. *Journal of the American Statistical Association* 87, 464-469.

Baringhaus, L., Henze, N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function. *Statistical & Probability Letters* 13, 269–274.

Carota, C. (2007). A family of power-divergence diagnostics for goodness-of-fit. *The Canadian Journal of Statistics*, 35(4), 549-561.

Cressie N.A.C., Read T.R.C. (1984). Multinomial goodness-of-Fit tests. *Journal of Royal Statistical Society Serious B*, 46(3) 440-464.

Cressie N.A.C., Read T.R.C., 1988. Goodness-of-fit statistics for discrete multivariate data. Springer, New York.

Cressie N.A.C., Read T.R.C. (1989). Pearson's X^2 and the Loglikelihood Ratio Statistic G^2: A Comparative Review. *International Statistical Review*, 57(1), 19–43.

Fisher, R.A. (1941). The negative binomial distribution. *Annals of Eugenics* 11, 182-187.

Freeman, M.F., Tukey, J.W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics* 21, 607-611.

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

Gurtler, N., Henze, N. (2000). Recent and classical goodness-of-fit tests for the Poisson distribution. *Journal of Statistical Planning and Inference* 90, 207-225.

Henze, N., Klar, B. (1995). Bootstrap based goodness of fit tests for the generalized Poisson model. *Communications in Statistics-Theory and Methods*, 24(7) 1875-1896.

Ivan, J.N., Wang, C., Bernardo, N.R. (2000). Explaining two-lane highway crash rates using land use and hourly exposure. *Accident Analysis & Prevention*, 32(6) 787-795.

Joshua, S.C., Garber, N.J. (1990). Estimating truck accident rate and involvements using linear and Poisson regression models. *Transportation Planning and Technology* 15, 41- 58.

Kim, B.S., Park, C. (1992). Some remarks on testing goodness of fit for the Poisson assumption. *Communications in Statistics-Theory and Methods* 21, 979–995.

Kocherlakota, S., Kocherlakota, K. (1986). Goodness of fit tests for discrete distributions. *Communications in Statistics-Theory and Methods* 15, 815–829.

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

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

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

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

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

Maycock, G., Hall, R.D. (1984). Accidents at 4-arm roundabouts. *Laboratory Report LR 1120*. Crowthorne, Berks, U.K. Transport Research Laboratory.

McCullagh, P., Nelder, J.A. (1989). *"Generalized linear models"*. 2nd Edition, Chapman & Hall, New York.

Miaou, S-P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus Negative Binomial regressions. *Proceedings of* the *73rd Annual Meeting of Transportation Research Board*, Washington D.C.

Miaou, S-P., Hu, P.S., Wright, T., Rathi, A.K., Davis, S.C. (1992). Relationships between truck accidents and highway geometric design: a Poisson regression approach. *Transportation Research Record* 1376, 10-18.

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

Miaou, S-P., Lum, H. (1993). Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis & Prevention*, 25(6) 689-709.

Mitra, S., Washington, S. (2007). On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis & Prevention, 39(3) 459-468.

Myers, R.H., Montgomery, D.C., Vining, G.G. (2002). *"Generalized linear models, with applications in engineering and the sciences"*. John Wiley & Sons.

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

Neyman, J. (1949). Contribution to the theory of the χ^2 Test. Proceedings of the First Berkeley *Symposium on Mathematical Statistics and Probability*, 239-273.

Read, T.R.C., Cressie, N.A.C. (1988). *"Goodness-of-fit statistics for discrete multivariate data"*. Springer-Verlag, New York.

SAS Institute Inc. (1999). The GENMOD procedure. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#genmod_toc.htm (Mar. 21, 2011).

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2) 461-464.

Spinelli, J.J., Stephens, M.A. 1997. Carmer-von Mises tests of fit for the Poisson distribution. *The Canadian Journal of Statistics*, 25(2) 257-268.

Sukhatme, P.V. (1938). On the Distribution of χ^2 in samples of the Poisson series. *Journal of the Royal Statistical Society: Serious B*, 5, 75–79.

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

Zhang, Y., Ye, Z., Lord, D. (2007). Estimating the dispersion parameter of the negative binomial distribution for analyzing crash data using a Bootstrapped maximum likelihood method. *Transportation Research Record* 2019, 15-21.