Accident Analysis and Prevention 59 (2013) 497–505

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Evaluating the double Poisson generalized linear model

Yaotian Zou[a,1], Srinivas Reddy Geedipally[b,*], Dominique Lord[c,2]

[a] School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907-2051, United States
[b] Texas Transportation Institute, Texas A&M University System, 110N, Davis Dr. 101, Arlington, TX 76013, United States
[c] Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, United States

ARTICLE INFO

Article history:
Received 29 August 2012
Received in revised form 9 July 2013
Accepted 12 July 2013

Keywords:
Double Poisson
Normalizing constant
Conway–Maxwell–Poisson
Gamma model
Generalized linear model

ABSTRACT

The objectives of this study are to: (1) examine the applicability of the double Poisson (DP) generalized linear model (GLM) for analyzing motor vehicle crash data characterized by over- and under-dispersion and (2) compare the performance of the DP GLM with the Conway–Maxwell–Poisson (COM-Poisson) GLM in terms of goodness-of-fit and theoretical soundness. The DP distribution has seldom been investigated and applied since its first introduction two decades ago. The hurdle for applying the DP is related to its normalizing constant (or multiplicative constant) which is not available in closed form. This study proposed a new method to approximate the normalizing constant of the DP with high accuracy and reliability. The DP GLM and COM-Poisson GLM were developed using two observed over-dispersed datasets and one observed under-dispersed dataset. The modeling results indicate that the DP GLM with its normalizing constant approximated by the new method can handle crash data characterized by over- and under-dispersion. Its performance is comparable to the COM-Poisson GLM in terms of goodness-of-fit (GOF), although COM-Poisson GLM provides a slightly better fit. For the over-dispersed data, the DP GLM performs similar to the NB GLM. Considering the fact that the DP GLM can be easily estimated with inexpensive computation and that it is simpler to interpret coefficients, it offers a flexible and efficient alternative for researchers to model count data.

1. Introduction

Due to the limited access to driving information, it is very difficult to identify factors (e.g., brake reaction time, inattention, etc.) directly influencing the number and severity of motor vehicle crashes in traffic safety analysis. Thus, instead of focusing on individual driver's information, most researchers approach crash causal or correlation analyses from a long term statistical view. In this regards, researchers try to associate the factors of interest with the frequency of crashes that occurs in a given space (roadway or intersection) and time period (Lord and Mannering, 2010). Therefore, statistical models have been the analysis tool of choice for analyzing the relationship between motor vehicle crashes and factors such as road section geometric design, traffic flow, weather, etc.

The traditional poisson distribution has been commonly used to model motor vehicle crashes. Despite its simple probabilistic structure, the traditional Poisson distribution has a strict assumption in that its single parameter does not allow for the flexibility of variance varying independently of the mean, which is often violated by its application to the over-dispersed (i.e., the sample variance is larger than the sample mean) and under-dispersed (i.e., the sample variance is smaller than the sample mean) crash data. Under-dispersion occurs on rare occasions and this often happens when the sample mean value is low (Lord and Mannering, 2010). Over-dispersed and under-dispersed data can lead to inconsistent standard errors of parameter estimates when the Poisson model is used (Cameron and Trivedi, 1998; Park and Lord, 2007).

Over the last three decades, the negative binomial distribution or model (NB or Poisson-gamma) has been quite popular to handle over-dispersed datasets. Although the mean-variance relationship of the NB is simple to manipulate so as to capture over-dispersion (Hauer, 1997), it has been found to have difficulties in handling the data characterized by under-dispersion (Lord et al., 2008a). Though crash data do not exhibit under-dispersion very often, it is observed more frequently in other fields of research (for example, see Guikema and Coffelt, 2008; Sellers and Shmueli, 2010; Borle et al., 2006). In order to manage data characterized by over-dispersion and under-dispersion, researchers have proposed alternative models such as the weighed Poisson distribution (Castillo and Perez-Casany, 2005), the generalized Poisson distribution (Consul, 1989) and the gamma count distribution (Winkelmann, 2008). However, these models suffered from important theoretical or logical soundness. For instance, the

* Corresponding author. Tel.: +1 817 462 0519.
E-mail addresses: zou21@purdue.edu (Y. Zou), srinivas-g@ttimail.tamu.edu (S.R. Geedipally), d-lord@tamu.edu (D. Lord).
[1] The work was done while the author was at Texas A&M University.
[2] Tel.: +1 979 458 3949.

0001-4575/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2013.07.017

weighted Poisson distribution has convergence restrictions and it is necessary to choose an appropriate function, which is sometimes difficult to do. With the generalized Poisson model, the bounded dispersion parameter when under-dispersion occurs greatly diminishes its applicability to count data (Famoye, 1993) and the model does not include the Poisson family in the interior of the parameter space (Castillo and Perez-Casany, 2005). The gamma count distribution assumes that observations are not independent where the observation for time $t-1$ would affect the observation for time t (Winkelmann, 2008; Cameron and Trivedi, 1998). This would become unrealistic if the time gap between the two observations is large, which can be problematic for analyzing crash data.

Among the distributions or models that have been documented in the literature, two distributions that can handle both under- and over-dispersion are particularly noteworthy. One is the Conway–Maxwell–Poisson (COM-Poisson or CMP) distribution (Conway and Maxwell, 1962; Shmueli et al., 2005; Kadane et al., 2006) and the other is the double Poisson (DP) distribution (Efron, 1986). Albeit first introduced in 1962, the statistical properties of the COM-Poisson have not been extensively investigated until recently. The COM-Poisson distribution and its generalized linear model (GLM) have been found to be very flexible to handle count data (Guikema and Coffelt, 2008; Lord et al., 2008a; Sellers et al., 2012; Francis et al., 2012). As for the DP, its distribution has seldom been investigated and applied since its first introduction about 25 years ago. A handful of research studies have mentioned that the hurdle for applying the DP is its normalizing constant (or multiplicative constant), which is not available in the closed form (Winkelmann, 2008; Hilbe, 2011; Zhu, 2012). They found that the results of the DP with its normalizing constant approximated by Efron's original method are not exact. Instead of using Efron's approximation method, this study documents a different method for handling the normalizing constant.

The objectives of this study are to: (1) examine the applicability of the DP GLM for analyzing motor vehicle crash data characterized by over- and under-dispersion and (2) compare the performance of the DP GLM with COM-Poisson GLM in terms of goodness-of-fit and theoretical soundness. Two empirical over-dispersed datasets (one for the high mean scenario and one for the low mean scenario) and one empirical under-dispersed dataset were used.

2. Background

This section describes the characterization and GLM framework of the DP and COM-Poisson models.

2.1. Double Poisson model

Based on the double exponential family, Efron (1986) proposed the double Poisson distribution. The double Poisson model, based on the distribution, has two parameters μ and θ. The approximate probability mass function (PMF) is given as:

$$P(Y=y)=f_{\mu,\theta}(y)=(\theta^{1/2}e^{-\theta\mu})\left(\frac{e^{-y}y^{y}}{y!}\right)\left(\frac{e\mu}{y}\right)^{\theta y},\ y=0,1,2,\ldots, \tag{1}$$

The exact double Poisson density is given as:

$$P(Y=y)=\tilde{f}_{\mu,\theta}(y)=c(\mu,\theta)f_{\mu,\theta}(y) \tag{2}$$

where the factor $c(\mu,\theta)$ can be calculated as:

$$\frac{1}{c(\mu,\theta)}=\sum_{y=0}^{\infty}f_{\mu,\theta}(y)\approx 1+\frac{1-\theta}{12\mu\theta}\left(1+\frac{1}{\mu\theta}\right) \tag{3}$$

With $c(\mu,\theta)$ being the normalizing constant nearly equal to 1. The constant $c(\mu,\theta)$ ensures that the density sums to unity.

The expected value and the standard deviation (SD) referring to the exact density $\tilde{f}_{\mu,\theta}(y)$ are estimated as follows:

$$E(Y)\approx\mu, \tag{4}$$

$$\mathrm{SD}(Y)\approx\left(\frac{\mu}{\theta}\right)^{1/2} \tag{5}$$

Thus, the double Poisson model allows for both over-dispersion ($\theta<1$) and under-dispersion ($\theta>1$). When $\theta=1$, the double Poisson distribution collapses to the Poisson distribution.

Particular focus should be given to the use of the normalizing constant which includes an infinite series $\sum_{y=0}^{\infty}f_{\mu,\theta}(y)$. Although Efron (1986) demonstrated the closed form approximation to the infinite series in Eq. (3) and even the approximate density in Eq. (2) are reasonably good for the case $\mu=10$, they are highly unreliable when μ is small (i.e., the sample mean is small). For instance, when $\mu=0.1$ and $\theta=1.5$, the closed form approximation solution turns into a negative value $(1+((1-\theta)/12\mu\theta)(1+(1/\mu\theta))=-1.13)$ and the sum of the approximate densities is not fairly close to unity $\left(\sum_{y=0}^{\infty}f_{\mu,\theta}(y)=1.11\right)$. Winkelmann (2008) and Hilbe (2011) also indicated that the normalizing constant is the hurdle in applying the DP. They found that the results of the DP with its normalizing constant approximated by Efron's original method are not exact.

Some researchers approached the DP model by completely ignoring the normalizing constant. However, this method is associated with both theoretical and practical limitations. It should be noted that the removal of the normalizing constant in the PMF will make the sum of all the likelihoods not equal to unity, which substantially diminishes the DP's mathematical appeal. Corresponding application based on this method also results in unreliable estimates. Zhu (2012) recently tested the DP without the normalizing constant, and found that the model provides a good fit for the mean, but does a terrible job for adequately capturing the variance.

In light of the inadequacy of the approximate density function and the aforementioned approximation to the normalizing constant in handling low mean data, it is important to approximate the normalizing constant with high accuracy and reliability. Given the fact that the infinite series $\sum_{y=0}^{\infty}f_{\mu,\theta}(y)$ is similar to the Poisson sum and that it converges very quickly especially when μ is small, this study takes the kth partial sum of the infinite series (i.e., the sum of the first k terms) to approximate its sum. This method makes it possible to compute the normalizing constant to some modest accuracy by adjusting the number k according to user's preference. After multiple trials on the selection of k values, we recommend the value of k to be as least twice as large as the sample mean.

For the DP GLM, the expected number of crashes per year is linked to the explanatory variables x_j by the following link function (similar to the traditional Poisson):

$$E(Y)\approx\mu=\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_j\right) \tag{6}$$

where the vector β_j is the coefficients to be estimated.

2.2. Conway–Maxwell–Poisson model

In order to model queues and service rates, Conway and Maxwell (1962) first introduced the COM-Poisson distribution as a generation of the Poisson distribution. However, this distribution was not widely used until Shmueli et al. (2005) further examined its statistical and probabilistic properties. Kadane et al. (2006) developed the conjugate distributions for the parameters of the COM-Poisson distribution.

The COM-Poisson distribution has two parameters with λ as the centering parameter and ν as the dispersion parameter. When ν is