Analytic Methods in Accident Research 5-6 (2015) 1–16

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in Negative Binomial models

Yajie Zou [a,1], Lingtao Wu [b,*], Dominique Lord [b,2]

[a] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195-2700, United States
[b] Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, United States

ARTICLE INFO

Article history:
Received 25 August 2014
Received in revised form
22 December 2014
Accepted 22 December 2014
Available online 12 March 2015

Keywords:
Sichel
Negative Binomial
Dispersion parameter
Traffic crashes
Empirical Bayes

ABSTRACT

Despite many statistical models that have been proposed for modeling motor vehicle crashes, the most commonly used statistical tool remains the Negative Binomial (NB) model. Crash data collected for safety studies may exhibit over-dispersion and a long tail (i.e., a few sites have unusually high number of crashes). However, some studies have shown that NB models cannot handle over-dispersed count data with a long tail adequately. So far, no work has investigated the performance of the dispersion parameter of the NB model when analyzing over-dispersed crash data with a long tail. The dispersion parameter of the NB model plays an important role in various types of transportation safety analysis. The first objective of this study is to examine whether the dispersion parameter can truly reflect the level of dispersion in over-dispersed crash data with a long tail. The second objective is to determine whether the dispersion term of the Sichel (SI) model can be used as an alternative to the dispersion parameter of the NB model. To accomplish the objectives of this study, crash data sets are simulated from NB and SI regression models using different values describing the mean and the dispersion level. For the simulated data sets, the dispersion parameter and dispersion term are estimated and compared to the true values. To complement the output of the simulation study, crash data collected in Texas are also used to compare the dispersion parameter and dispersion term. The results from this study suggest that the dispersion parameter of the NB model can erroneously estimate the level of dispersion in over-dispersed count data with a long tail and the dispersion term of the SI model is more reliable in estimating the true level of dispersion. Thus, considering the findings in this study, it is believed that the dispersion term may offer a viable alternative for analyzing over-dispersed crash data with a long tail.

* Corresponding author. Tel.: +1 979 587 3518; fax: +1 979 845 6481.
E-mail addresses: zouyajie@gmail.com (Y. Zou), wulingtao@gmail.com (L. Wu), d-lord@tamu.edu (D. Lord).
[1] Tel.: +1 936 245 5628; fax: +1 206 543 5965.
[2] Tel.: +1 979 458 3949; fax: +1 979 845 6481.

http://dx.doi.org/10.1016/j.amar.2014.12.002
2213-6657/© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

From a statistical point of view, the occurrence of highway crashes can be treated as random events by assuming that there is an underlying mean crash rate for each individual site (Park et al., 2010). What makes the analysis difficult in modeling crash data is that the crash data are often found to exhibit over-dispersion, meaning that the variance is greater than the mean (Park and Lord, 2009). Lord et al. (2005) provided a fundamental definition that the over-dispersion arises from the actual nature of the crash process. To accommodate over-dispersion in crash data, many mixed-Poisson models have been proposed by transportation safety analysts, such as the Negative Binomial (NB, also known as Poisson-gamma) models (Poch and Mannering, 1996; Miaou and Lord, 2003), zero-inflated models (Shankar et al., 1997), the Poisson-lognormal (Aguero-Valverde and Jovanis, 2008), the Conway–Maxwell–Poisson (Lord et al., 2008b), the Poisson–Weibull (Cheng et al., 2013), etc. (for a comprehensive review of the mixed-Poisson models used in transportation safety analysis, see Mannering and Bhat (2014)). These statistical models are in fact used as an approximation for modeling crash data. Among these mixed-Poisson models, the NB model remains the most frequently used statistical model for accommodating the over-dispersion observed in the crash data (Lord and Mannering, 2010). Reasons for the popularity of the NB models include: (1) the NB model provides a simple way to manipulate the relationship between the mean and the variance (Lord and Mannering, 2010); (2) the dispersion parameter of the NB model plays an important role in transportation safety analysis. Besides the mixed-Poisson models, the random parameters count models (Anastasopoulos and Mannering, 2009; Chen and Tarko, 2014), finite mixture and Markov switching models (Park and Lord, 2009; Malyshkina et al., 2009; Zou et al., 2013b, 2014), generalized ordered-response models (Castro et al., 2012; Bhat et al., 2014) and quantile regression models (Qin and Reyes, 2011) have been proposed for analyzing the crash-frequency data.

The dispersion parameter of the NB model is critical for estimating the weight factor of the empirical Bayes (EB) method (Hauer et al., 1988; Hauer, 1997) and for building confidence intervals for evaluating and screening highway projects (Wood, 2005). Since the above two types of analysis are commonly used in highway safety, it is necessary to obtain reliable estimates of the dispersion parameter. It has been shown that the low sample mean and small sample size can significantly influence the estimation of the dispersion parameter of NB models using the maximum likelihood estimation method and Bayesian method (Maher and Summersgill, 1996; Lord, 2006; Lord and Miranda-Moreno, 2008). To avoid or minimize an unreliably estimated dispersion parameter, Lord (2006) also summarized the minimum sample size for different sample means.

For NB models, the gamma distribution assumed in the probabilistic error term related to the mean of the Poisson variable can be restrictive in terms of its ability to account for heterogeneity across observations (Park et al., 2010). For example, Guo and Trivedi (2002) have reported that NB regression models have difficulties modeling heavily over-dispersed data with a long-tail and relatively high mean value because a negligible probability is usually assigned to high counts. Recently, the Sichel distribution (SI, also known as the Poisson-generalized inverse Gaussian distribution) has been introduced by Zou et al. (2013a) for calculating EB estimates. The SI distribution is a compound Poisson distribution, which mixes the Poisson distribution with the generalized inverse Gaussian distribution. Previous studies (Stein et al., 1987; Gupta and Ong, 2005) have shown that the SI distribution is useful as a model for over-dispersed count data with a long tail. Among different mixed-Poisson models, it is found that the NB and SI models both have the quadratic variance–mean relationship. Similar to the dispersion parameter of the NB model, a dispersion term of the SI model can be defined to measure the level of dispersion in the data. This dispersion term can be easily used by transportation safety analysts to obtain reliable EB estimates within the SI modeling framework (Zou et al., 2013a).

Considering the importance of the dispersion parameter of the NB model in transportation safety analysis, the objective of this study is to examine whether or not the traditionally used dispersion parameter can truly reflect the level of dispersion in over-dispersed crash data with a long tail and whether the dispersion term of the SI model can be used as an alternative to the dispersion parameter. To accomplish the objectives of this study, crash data sets are simulated from NB and SI models using different combinations of fixed regression parameters describing the mean and the dispersion level. For the simulated datasets, the dispersion parameter and dispersion term are estimated and compared to the true values. The simulation analysis is carried out in this study for the following reason: when analyzing real crash data, the true values of regression parameters and the dispersion level of the crash data are seldom known in practice. In contrast, in a simulation, it is possible to generate crash data with known regression parameters and dispersion levels. The simulation analysis was used in previous transportation safety studies (Lord, 2006; Francis et al., 2012) to characterize the performance of different estimators. To complement the output of the simulation study, crash data collected in Texas are also used to compare the dispersion parameter and dispersion term.

2. Methodology

The NB models have the following probabilistic structure: the number of crashes Y_{it}, at the ith site and time period t, when conditional on its mean μ_{it} is Poisson distributed and independent over all sites and time periods (Miaou and Lord, 2003):

$$Y_{it}|\mu_{it} \sim \text{Poisson}(\mu_{it}), \quad i=1,2,\ldots,I \quad \text{and} \quad t=1,2,\ldots,T \tag{1}$$