Accident Analysis and Prevention 91 (2016) 10–18

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data

Mohammadali Shirazi [a], Dominique Lord [a,*], Soma Sekhar Dhavala [b], Srinivas Reddy Geedipally [c]

[a] Zachry Department of Civil Engineering, Texas A&M University, College Station, TX 77843, United States
[b] Perceptron Learning Solutions Pvt Ltd, Bengaluru, India
[c] Texas A&M Transportation Institute, Texas A&M University, College Station, TX 77843, United States

ARTICLE INFO

Article history:
Received 9 November 2015
Received in revised form 21 February 2016
Accepted 22 February 2016
Available online 3 March 2016

Keywords:
Negative binomial
Dirichlet process
Generalized linear model
Crash data

ABSTRACT

Crash data can often be characterized by over-dispersion, heavy (long) tail and many observations with the value zero. Over the last few years, a small number of researchers have started developing and applying novel and innovative multi-parameter models to analyze such data. These multi-parameter models have been proposed for overcoming the limitations of the traditional negative binomial (NB) model, which cannot handle this kind of data efficiently. The research documented in this paper continues the work related to multi-parameter models. The objective of this paper is to document the development and application of a flexible NB generalized linear model with randomly distributed mixed effects characterized by the Dirichlet process (NB-DP) to model crash data. The objective of the study was accomplished using two datasets. The new model was compared to the NB and the recently introduced model based on the mixture of the NB and Lindley (NB-L) distributions. Overall, the research study shows that the NB-DP model offers a better performance than the NB model once data are over-dispersed and have a heavy tail. The NB-DP performed better than the NB-L when the dataset has a heavy tail, but a smaller percentage of zeros. However, both models performed similarly when the dataset contained a large amount of zeros. In addition to a greater flexibility, the NB-DP provides a clustering by-product that allows the safety analyst to better understand the characteristics of the data, such as the identification of outliers and sources of dispersion.

1. Introduction

Regression models have different applications in highway safety. They can be used for predicting the number of crashes, evaluating roadway safety, screening variables and identifying hazardous sites. As documented in Lord and Mannering (2010) and more recently in Mannering and Bhat (2014), extensive research studies have been devoted to develop innovative and novel statistical models to estimate or predict the number of crashes and evaluate roadway safety. These statistical models specifically deal with unique characteristics that are associated with crash data. As such, crash data can often be characterized with over-dispersion, heavy tail and many observations with the value zero. These unique characteristics inspired a few researchers to propose several new models that aimed at overcoming the limitations associated with the most commonly used model in highway safety literature, the negative binomial (NB) model (also known as the Poisson-gamma model).

Recent research has shown that the NB model can be significantly affected by datasets characterized by a heavy tail (Zou et al., 2015). According to Guo and Trivedi (2002), the NB regression model cannot properly capture the heavy tail because a negligible probability is assigned to large counts. Heavy tails can be caused by the data generating process itself (i.e., including observations with very large counts), or they can also be attributed to datasets that have excess zeros. In the latter case, the heavy tails are created by shifting the overall sample mean closer to zero, which increases the spread of the observations (Lord and Geedipally, 2016). Over the last two or three years, a new series of multi-parameter models

* Corresponding author.
E-mail addresses: alishirazi@tamu.edu (M. Shirazi), d-lord@tamu.edu (D. Lord), soma.dhavala@gmail.com (S.S. Dhavala), srinivas-g@tti.tamu.edu (S.R. Geedipally).

http://dx.doi.org/10.1016/j.aap.2016.02.020
0001-4575/

(i.e., models with several shape/scale parameters) that mixes the NB distribution with other distributions have been developed for analyzing such datasets. The NB-Lindley (NB-L) model (Geedipally et al., 2012) and the NB-generalized exponential (NB-GE) (Vangala et al., 2015) are examples of such types of models. This paper continues describing research in this line of modeling work.

A recurring theme in many multi-parameter models is to consider a mixing distribution at the heart of the generative model. For example, one can see the NB as a Poisson-gamma mixture or the NB-L as a mixture of the NB and the Lindley distributions (note: the Lindley distribution itself is a mixture of two Gamma distributions). There are primarily three major ingredients to eliciting such mixtures, which offer a greater degree of flexibility in model construction:

1 The mixing weights: the mixing weights determine the relative weight of the individual mixing components.
2 The shape and characteristics of the mixing components or the constituent members of the mixtures, and
3 The level: in the context of hierarchical/multi-level modeling, at which level the mixture distribution is elicited.

A transportation safety analyst might have a preference to choose or rather not to choose a particular mixture. In all cases, the analyst is required to make certain assertions about the mixture components. One way to retain the modeling flexibility and yet not be overly concerned about the assertions is to express the uncertainty explicitly by considering a random mixing distribution. The Dirichlet process (DP), a widely used prior in Bayesian nonparametric literature, allows such representation (Antoniak, 1974; Escobar and West, 1995). One way to think about the DP is as an infinite mixture distribution, where the number of unique components and the component characteristics themselves can be learned from the data.

There has been a phenomenal growth in theory, inference and applications concerning the DP and its related processes in the last decade; recent monographs on Bayesian nonparametric devoting significant portion on the DP and related processes is a testimony to that effect (Hjort et al., 2010; Mitra and Muller, 2015). On the application side, the DP has been applied in numerous fields ranging from network modeling (Ghosh et al., 2010) to Bioinformatics (Dhavala et al., 2010; Argiento et al., 2015) to Psychometrics (Miyazaki and Hoshino, 2009) to name a few. In particular, the application of the DP to account for over-dispersion in count data has been considered in Mukhopadhyay and Gelfand (1997) and Carota and Parmigiani (2002), with Binomial and Poisson based likelihoods. More details about the DP, its structure and computational details are discussed in Sections 2 and 6.

The objective of this study is to develop and document a new method to model over-dispersed data with a heavy tail. The model is introduced based on the Bayesian hierarchical modeling framework as a mixture of the NB distribution and a random distribution characterized by the DP. The proposed model can be motivated, first, by looking at the NB model as a mixture of the Poisson and Gamma distributions. As an extension of the Poisson model, the Poisson-gamma was developed assuming that the Poisson parameter is measured with a random error; this random error itself is gamma distributed. The Poisson-gamma mixture is thought to be a better alternative to accommodate possible over-dispersion in data (Hilbe, 2011). Second, it can be motivated by looking at the NB-L model as a mixture of the negative binomial and the Lindley distributions. The NB-L model can overcome the NB limitations when data are over-dispersed and have many zeros. Essentially, as discussed above, although mixture models are providing better alternatives, they assume the shape and density of the distributions to be fixed. However, we can obtain even more flexibility by assuming that the mixing distribution itself is random. Given this motivation in mind, the current research plans to develop a model as a mixture of the negative binomial and a random distribution characterized by the DP.

In addition to providing greater flexibility, the proposed model groups crash data into a finite number of clusters as one of its by-products. The clustering property of the mixture model can lend insights to learn more about the domain or the data. This can be used to (1) identify outliers; (2) study the sites that fall into the same clusters to identify the safety issues and get insights to implement appropriate countermeasures; and (3) examine sources of dispersions (Peng et al., 2014).

2. Characteristics of the Dirichlet process (DP)

Traditionally, the Bayesian parametric inference mechanism considers a parametric distribution $F_0(.|\theta)$, where θ is a finite vector of parameters, as a prior for the unknown parameter. However, constraining the model within specific parametric families could limit the scope of the inference. To overcome this difficulty, in context of the Bayesian nonparametric (or semiparametric) modeling, a random prior distribution is considered for the parameter as opposed to choosing a prior distribution from a known parametric family. The prior is placed over infinite-dimension space of distribution functions. In that sense, it gives more flexibility to the parameter inference mechanism by providing a wide range of prior distributions.

The DP (Ferguson, 1973, 1974) is a stochastic process that is usually used as a prior in Bayesian nonparametric (or semiparametric) modeling. Escobar and West (1998) define the DP as a random probability measure over the space of all probability measures. In that sense, the DP is considered as a distribution over all possible distributions; that is, each draw from the DP is itself a distribution. Below, we provide a formal definition and characterization of the DP. For a gentle introduction and motivation to DP as an extension of the finite dimensional mixtures to infinite dimensional, the interested readers are referred to Teh (2010) and Gelman et al. (2014).

Let $A_1, A_2, .., A_r$ be any finite measurable partitions of the parameter space (Θ). Let us assume τ be a positive real number and $F_0(.|\theta)$ be a continuous distribution over Θ. Then, $F(.) \sim DP\left(\tau, F_0(.|\theta)\right)$ if and only if (Escobar and West, 1998):

$$\left(F(A_1), F(A_2), \ldots, F(A_r)\right) \sim$$
$$Dirichlet\left(\tau F_0\left(A_1|\theta\right), \tau F_0\left(A_2|\theta\right), \ldots, \tau F_0\left(A_r|\theta\right)\right) \quad (1)$$

where τ is defined as the precision (or concentration) parameter and $F_0(.|\theta)$ as the base (or baseline) distribution. Note that based on the Dirichlet distribution properties, for each partition $A \subset \Theta$, we have:

$$E(F(A)) = F_0(A|\theta)$$

$$var(F(A)) = \frac{F_0\left(A|\theta\right)\left(1 - F_0\left(A|\theta\right)\right)}{1+\tau}$$

Therefore, the base distribution $F_0(.|\theta)$ and the precision parameter τ play significant roles in the DP definition. The expectation of the random distribution $F(.)$ is the base distribution $F_0\left(.|\theta\right)$. Likewise, the precision parameter τ controls the variance of the random distribution around its mean. In other words, τ measures the variability of the target distribution around the base distribution. As $\tau \to \infty$, we would have $F(.) \to F_0(.|\theta)$ while, on the other hand, as $\tau \to 0$, the random distribution $F(.)$ would deviate further away from $F_0(.|\theta)$.

Eq. (1) defines the DP indirectly through the marginal probabilities assigned to finite number of partitions. Therefore, it gives no intuition on realizations of $F(.) \sim DP\left(\tau, F_0(.|\theta)\right)$. To simulate