Accident Analysis and Prevention 50 (2013) 289–297

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates

Jonathan Aguero-Valverde*

Programa de Investigación en Desarrollo Urbano Sostenible, Universidad de Costa Rica, Barrio Los Profesores, Calle B, No 11, Mercedes, San Pedro, San José 11503, Costa Rica

ARTICLE INFO

Article history:
Received 1 February 2012
Received in revised form 13 April 2012
Accepted 30 April 2012

Keywords:
Full Bayes
Zero inflated models
Random effects
Ranking of sites

ABSTRACT

In recent years, complex statistical modeling approaches have being proposed to handle the unobserved heterogeneity and the excess of zeros frequently found in crash data, including random effects and zero inflated models. This research compares random effects, zero inflated, and zero inflated random effects models using a full Bayes hierarchical approach. The models are compared not just in terms of goodness-of-fit measures but also in terms of precision of posterior crash frequency estimates since the precision of these estimates is vital for ranking of sites for engineering improvement. Fixed-over-time random effects models are also compared to independent-over-time random effects models.

For the crash dataset being analyzed, it was found that once the random effects are included in the zero inflated models, the probability of being in the zero state is drastically reduced, and the zero inflated models degenerate to their non zero inflated counterparts. Also by fixing the random effects over time the fit of the models and the precision of the crash frequency estimates are significantly increased.

It was found that the rankings of the fixed-over-time random effects models are very consistent among them. In addition, the results show that by fixing the random effects over time, the standard errors of the crash frequency estimates are significantly reduced for the majority of the segments on the top of the ranking.

1. Introduction

One of the most important tasks of highway safety practitioners is the identification of locations in need of engineering improvements to reduce the number of crashes. The literature on the subject of identifying those sites includes a series of papers by Hauer (1996) and Hauer et al. (2002, 2004). Most ranking techniques rely on the estimation of the crash frequency for each segment or a function of it, such as the difference between the expected crash frequency in the site and the expected crash frequency in similar sites (Hauer, 1996; Tarko and Kanodia, 2004; Aguero-Valverde and Jovanis, 2007, 2009).

For modeling the expected crash frequency as well as better understanding the factors that affect the risk of a crash, researchers have used several count models of varying complexity, from Poisson to zero state Markov switching models. Jovanis and Chang (1986) presented one of the early uses of the Poisson regression for modeling crash frequencies. Other early adopters of Poisson regression include Jones et al. (1991) and Miaou and Lum (1993). Later, researchers adopted Poisson gamma (also known as negative binomial) models to account for the overdispersion frequently found in crash data (Shankar et al., 1995; Persaud and Mucsi, 1995; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000; Donnell and Mason, 2006).

In recent years, more complex statistical modeling approaches have being proposed to handle the unobserved heterogeneity and the excess of zeros frequently found in crash data. Among them, random effects and random parameter models have gained popularity in the field. Random effects models were proposed to account for correlation among observations in the data (Lord and Mannering, 2010). Some applications in highway safety can be found in Johansson (1996), Shankar et al. (1998), MacNab (2004), Miaou and Lord (2003), Aguero-Valverde and Jovanis (2006) and Quddus (2008). More recently, random parameters models were proposed to control for unexplained heterogeneity in the data. Some examples of this approach are Anastasopoulos and Mannering (2009) and El-Basyounny and Sayed (2009).

Zero-inflated (ZI) models have being proposed to account for the excess of zeros, compared to the number of zeros expected under a Poisson or Poisson-gamma model, exhibited by many crash datasets. One of the first reported works to incorporate zero inflated models in road safety was the paper by Shankar et al. (1997). Other examples of zero inflated models in highway safety are Carson and Mannering (2001), Shankar et al. (2004), and Qin et al. (2005). These

* Tel.: +506 22834927; fax: +506 22834815.
E-mail address: jaguero@produs.ucr.ac.cr

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.04.019

models operate under the assumption of two states existing for the data: the "zero" state and the normal count state. This fact has led some researchers to criticize the application of these models in highway safety (Lord et al., 2005, 2007), since the segments in the zero state have a long term mean equal to zero implying that those segments are totally safe.

To overcome the issue of the long term mean equal to zero Malyshkina and Mannering (2010) proposed a Markov switching model that allows individual segments to change states over time. This model was implemented in a full Bayes approach since the likelihood function does not have a closed form; therefore, maximum likelihood estimation was not feasible. Furthermore, full Bayes estimation allows for direct statistical estimation of the road segment state.

Using also a full Bayes approach, this work proposes models that do not allow segments change states over time but concentrates on including random effects in the specification. Random effects, zero inflated, and zero inflated models with random effects are estimated. The models are compared not just in terms of goodness-of-fit measures but also in terms of precision of posterior crash frequency estimates since the precision of these estimates is vital for ranking of sites for engineering improvement. Other objective is to compare models with and without random effects fixed over time.

Even though, previous studies have analyzed and compared zero-inflated and Poisson or Poisson-gamma models for crash frequency in terms of goodness-of-fit, the effect of zero inflated models in the precision of crash frequency estimates is unknown. Ideally, one would try to assess the accuracy rather than precision of those estimates, but the "real" value of crash frequency is unknown. As an alternative, one can use simulated data for comparing the models so that the "real" value of crash frequency is known but this would defeat the purpose of comparing the models; the statistical model whose distribution more closely resembles the probability distribution used to simulate the dataset will always have a better fit. To circumvent this issue one can use the precision of the estimates. Estimates whose standard deviation is lower (higher precision) are statistically better.

In the present formulation, the segment state for each year is independent of the other years; therefore, it allows changes in state between years. Furthermore, all the parameters in Bayesian statistics are regarded as random in nature; hence, the expectation for the number of crashes cannot be equal to zero unless the probability of being in the zero state concentrates all its mass at one.

The fit of the models and the precision of the frequency estimates are analyzed for models where the random effects are fixed over time. Analog to the spatial correlation term (Aguero-Valverde and Jovanis, 2006, 2008, 2010), by using a fixed-over-time random effects, the segments estimates 'pool strength' from neighboring years improving model estimation. This is especially true in circumstances with high random variability in the data, such as is the case of most crash data, particularly when a high number of zeros is present.

Finally, the ranking of sites based on the posterior crash frequency estimates is compared for all proposed models in order to explore the effect that different modeling approaches have in the rankings. This paper is organized as follows: the next section describes the statistical methodology used; then, the dataset is described, followed by the discussion of results, conclusions and recommendations for future research.

2. Methodology

Full Bayes hierarchical approach is used for all the models estimated in this work. Markov chain Monte Carlo, also known as Markov chain simulation is employed to draw samples from the target posterior distribution of the parameters (Carlin and Louis, 1996). Details about Markov chain simulation are beyond the scope of this paper, interested readers can refer to Carlin and Louis (1996) and Gelman et al. (2003).

The formulation of the models is presented from the traditional Poisson-gamma to the more complex zero inflated models with random effects. Five count distributions are tested: Poisson gamma or negative binomial, Poisson lognormal, Poisson zero inflated, Poisson zero inflated lognormal and Poisson zero inflated gamma.

2.1. Poisson gamma

Poisson gamma or negative binomial (NB) is the most popular count distribution used in highway safety for crash frequency models. The model is specified as follows: at the first stage the crash counts are modeled as a Poisson process:

$$y_{it} \sim \text{Poisson} \quad (\theta_{it}) \tag{1}$$

where y_{it} are the observed number of crashes in segment i at time t (in years), and θ_{it} are the expected Poisson rate (i.e. the expected crash frequency) for segment i at time t.

The Poisson rate is modeled as a function of the covariates following the log-link shown in Eq. (2):

$$\log \quad (\theta_{it}) = \beta_0 + \sum_k \beta_k x_{itk} + \varepsilon_{it} \tag{2}$$

where β_0 is the intercept, β_k is the coefficient for covariate k, x_{itk} is the value for the kth covariate (or any suitable transformation of the covariate), for segment i, at time t and ε_{it} is the error term for segment i at time t.

At the second stage, the coefficients are modeled using non-informative Normal priors (i.e. $\beta_k \sim N(0, 1000)$) while the exponent of the error term or random effect is modeled as a gamma distribution:

$$e^{\varepsilon_{it}} \,|\, \phi \overset{iid}{\sim} \text{Gamma} \quad (\phi, \phi) \tag{3}$$

where ϕ controls the amount of extra-Poisson variation due to heterogeneity among segments. The gamma distribution of the error term has a mean of one and a variance of $1/\phi$. The parameter ϕ is also known as the dispersion parameter. A gamma (0.01, 0.01) was used as hyper-prior for ϕ. This hyper-prior was selected because it introduces little prior information into the model, reduces convergency times, and improves model identifiability.

The error term was also modeled as a random effect fixed over time. As in the case of the prior model, the exponent of the error term has a gamma prior distribution:

$$e^{\varepsilon_i} \,|\, \phi \overset{iid}{\sim} \text{Gamma} \quad (\phi, \phi) \tag{4}$$

2.2. Poisson lognormal

Poisson lognormal models have been used only recently in highway safety since the marginal distribution of this model does not have a closed form like the Poisson gamma model; therefore, they are typically implemented using the Bayesian approach. Some recent applications of these models in highway safety are: Miranda-Moreno et al. (2005), Aguero-Valverde and Jovanis (2006), Ma et al. (2007), and Park and Lord (2007). For this model the Poisson rate is modeled using a log-normal distribution:

$$\log \quad (\theta_{it}) = \beta_0 + \sum_k \beta_k x_{itk} + \nu_{it} \tag{5}$$