Analytic Methods in Accident Research 14 (2017) 1–9

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

A negative binomial crash sum model for time invariant heterogeneity in panel crash data: Some insights

Ghasak I.M.A. Mothafer [a], Toshiyuki Yamamoto [b], Venkataraman N. Shankar [c,*]

[a] Department of Civil Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
[b] Institute of Materials and Systems for Sustainability, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
[c] Department of Civil and Environmental Engineering, Penn State University, Pennsylvania, 226 C Sackett Building, PA 16802, USA

ARTICLE INFO

Article history:
Received 14 September 2016
Received in revised form 27 December 2016
Accepted 27 December 2016
Available online 19 January 2017

Keywords:
Time invariant heterogeneity
Types of crashes
Random effects
Random effects Poisson gamma (REPG)
Negative binomial (NB)

ABSTRACT

This paper presents a negative binomial crash sum model as an alternative for modeling time invariant heterogeneity in short panel crash data. Time invariant heterogeneity arising through multiple years of observation for each segment is viewed as a common unobserved effect at the segment level, and typically treated with panel models involving fixed or random effects. Random effects model unobserved heterogeneity through the error term, typically following a gamma or normal distribution. We take advantage of the fact that gamma heterogeneity in a multi-period Poisson count modeling framework is equivalent to a negative binomial distribution for a dependent variable which is the summation of crashes across years. The Poisson panel model referred to in this paper is the random effects Poisson gamma (REPG). In the REPG model, the dependent variable is an annual number of a specific crash type. The multi-year crash sum model is a negative binomial (NB) model that is based on three consecutive years of crash data (2005–2007). In the multi-year crash sum model, the dependent variable is the sum of crashes of a specific type for the three-year period. Four categories (in addition to total crashes) of crash types are considered in this study including rear end, sideswipe, fixed objects and all-other types. The empirical results show that when time effects are insignificant in short panels such as the one used in this study, the three-year crash sum model is a computationally simpler alternative to a panel model for modeling time invariant heterogeneity while imposing fewer data requirements such as annual measurements.

1. Introduction

In the recent years, numerous methodological advances for modeling count data have been considered (see Lord and Mannering, 2010 and Mannering and Bhat, 2014). The state-of-the-art can be classified into the following: a) random parameters in crash models (on random-parameters negative binomial model, see Anastasopoulos and Mannering, 2009; Venkataraman et al., 2014; Coruh et al., 2015; on random-parameters Poisson model, see Barua et al., 2016), b) multivariate/bivariate crash count models (Barua et al., 2016; Dong et al., 2016; Heydari et al., 2016; Serhiyenko et al., 2016), c) finite mixture, two-state Markov switching, neural network and latent class models (Malyshkina et al., 2009; Park and Lord, 2009; Behnood et al., 2014; Behnood and Mannering, 2016; Zeng et al., 2016), d) spatial/temporal correlation models (Chiou et al.,

* Corresponding author.
E-mail addresses: mothafer.g.i.m.a@c.mbox.nagoya-u.ac.jp (G.I.M.A. Mothafer), yamamoto@civil.nagoya-u.ac.jp (T. Yamamoto), shankarv@engr.psu.edu (V.N. Shankar).

http://dx.doi.org/10.1016/j.amar.2016.12.003
2213-6657/

2014; Chiou and Fu, 2015; Hong et al., 2016), e) zero-inflated crash count models (Shankar et al., 1997; Lord et al., 2005; Dong et al., 2014), and f) crash rate models (converting the crash-count variable into a continuous variable) (on Tobit model see Anastasopoulos, 2016; Anastasopoulos et al., 2016; Sarwar and Anastasopoulos, 2017). Our main contribution in this paper will belong to category d), precisely the temporal effect to add further insight into this regard. The motivation for our work is supported by recent work on the influence of unobserved heterogeneity (Russo et al., 2014; Behnood and Mannering, 2015; Anastasopoulos and Mannering, 2016; Norros et al., 2016).

In the classical approach of modeling the crash count data, it is customary to assume the data is cross-sectional in nature (see for example, Ivan et al., 2000; Lord, 2000; Lyon et al., 2003; Miaou and Lord, 2003; Lord et al., 2005; Miaou and Song, 2005; Ma et al., 2008). Alternatively, when time effects are considered, we have a combination of cross-sectional data and time series data (also known as the panel data) in which the duration of observations is included (Chin and Quddus, 2003; Kumara and Chin, 2004; Quddus, 2008; Law et al., 2009). Time effects shared across observations result in the efficiency problem (Shankar et al., 1995; Greene, 2003). Ignoring time invariant heterogeneity causes downward bias of standard errors, and results in the consequence that parameters that are insignificant in reality will be incorrectly identified as significant, and therefore included in the model. A recent study by Mannering et al. (2016) discusses the various form of unobserved heterogeneity and consequences for parameter estimation. Time invariant heterogeneity is identified as an important aspect of unobserved heterogeneity, and the study emphasizes the importance of further study on this subject.

Time invariant heterogeneity in crash data can be modeled by accounting for repeated observation effects through a negative multinomial density (Ulfarsson and Shankar, 2003). In this approach, the negative binomial density is modified to account for contributions from each time period. In the second approach, the error term across time observations is treated as a random effect (Shankar et al., 1998) that follows an arbitrary continuous distribution.[1] A more general approach is to treat the time effect via the year indicator as a random parameter in a random parameters count model. If some or all of the year indicators are random, then, it implies that the serial correlation effect is stochastic and not constant across years. The random effects model is a constrained version of this model, where the intercept alone is random. Sittikariya et al. (2005) proposed a zero inflated Poisson (ZIP) model to account for excess zeroes in the crash data. In their method, the authors used the negative multinomial approach to adjust the standard errors. By comparing the negative multinomial standard errors and cross sectional negative binomial standard errors, they used a loading factor principle which represented the level of inflation in standard errors of the parameter estimates due to serial correlation. Neither of the approaches or other published literature on serial correlation in count models addresses the impact of cumulative exposure. Under cumulative exposure, one can visualize the crash model to be composed of multiple years of observation, as opposed to the conventional one-year window. With multiple years of observation, the problem of excess zeros is potentially mitigated, while the problem of time series effects is now treated as a single cumulative effect as opposed to a common unobserved effect across time periods for any given segment.

The objective of this study is to evaluate the parameters of a count model under the influence of time invariant heterogeneity and compare the magnitudes and standard errors with those estimated from a cumulative crash count model. The cumulative crash count model, also termed here in this paper as the crash sum model would represent a single cross sectional analysis of crashes summed up across the entire time period.

2. Data

The crash count panel data used in the analysis is obtained from Washington State Department of Transportation crashes records for three years from 2005 to 2007. The panel data were collected from a multi-lane divided interstate highway (Interstate 5) in the State of Washington, USA. Table 1 shows the crash frequencies by type and year across 274 segments. Rear end, sideswipe and fixed object crash types exhibited the highest occurrence records among all the other types of crashes. Since the number of crashes that resulted from overturn, same direction, head-on and other miscellaneous types is limited, it was not plausible to statistically distinguish among these types. Hence, these categories were combined into a single category denoted as all-other (here and after). Tables 1 and 2 provide summaries of the crash, traffic volume and geometric data used in this study. These variables are described in detail in Mothafer et al. (2016).

3. Methodology

Modeling crash count panel data with random effects begins by assuming the error term α follows a distribution $f(\alpha)$ (for example gamma, or Gaussian distribution, Cameron and Trivedi, 2013). Let i ($i = 1, 2,\ldots,I$) be an index of crash types, t ($t = 1, 2,\ldots,T$) be an index representing the year of observation in the panel, and q ($q = 1, 2,\ldots,Q$) be an index of segment number, respectively. Then we can write the joint probability of the observed crash count variable y_{itq} for a given crash type i observed during time $t = 1, 2,\ldots,T$ on segment q as:

$$P[y_{i1q}, y_{i2q}, \ldots, y_{iTq}] = \int_0^\infty P[y_{i1q}, y_{i2q}, \ldots, y_{iTq}|\alpha_{iq}] f(\alpha_{iq}) d\alpha_{iq} = \int_0^\infty \left[\prod_{t=1}^{T} P[y_{itq}|\alpha_{iq}]\right] f(\alpha_{iq}) d\alpha_{iq} \quad (1)$$

[1] An alternative approach uses an arbitrary discrete distribution representation of unobserved heterogeneity, which generates a class of models called finite mixture models (Cameron and Trivedi, 2001).