Analytic Methods in Accident Research 2 (2014) 52–69

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

Crash frequency modeling using negative binomial models: An application of generalized estimating equation to longitudinal data

Mojtaba A. Mohammadi [a,*], V.A. Samaranayake [b,1], Ghulam H. Bham [c,2]

[a] Civil, Architecture, and Environmental Engineering, Missouri University of Science and Technology, 219 Butler-Carlton Hall, 1401 N. Pine St., Rolla, MO 65409, USA
[b] Mathematics and Statistics, Missouri University of Science and Technology, 202 Rolla Building, 400 W. 12th St., Rolla, MO 65409-0020, USA
[c] Civil Engineering, University of Alaska Anchorage, 3211 Providence Drive, Anchorage, AK 99508-4614, USA

ARTICLE INFO

Article history:
Received 11 September 2013
Received in revised form
6 July 2014
Accepted 6 July 2014
Available online 29 August 2014

Keywords:
Generalized estimation equation
Longitudinal analysis
Temporal correlation
Crash frequency model
Autocorrelation
Autoregressive

ABSTRACT

The prediction of crash frequency models can be improved when several years of crash data are utilized, instead of three to five years of data most commonly used in research. Crash data, however, generates multiple observations over the years that can be correlated. This temporal correlation affects the estimated coefficients and their variances in commonly used crash frequency models (such as negative binomial (NB), Poisson models, and their generalized forms). Despite the obvious temporal correlation of crashes, research analyses of such correlation have been limited and the consequences of this omission are not completely known. The objective of this study is to explore the effects of temporal correlation in crash frequency models at the highway segment level.

In this paper, a negative binomial model has been developed using a generalized estimating equation (GEE) procedure that incorporates the temporal correlations amongst yearly crash counts. The longitudinal model employs an autoregressive correlation structure to compare to the more traditional NB model, which uses a Maximum Likelihood Estimation (MLE) method that cannot accommodate temporal correlations. The GEE model with temporal correlation was found to be superior compared to the MLE model, as it does not underestimate the variance in the coefficient estimates, and it provides more accurate and less biased estimates. Furthermore, an autoregressive correlation structure was found to be an appropriate structure for longitudinal type of data used in this study. Ten years (2002–2011) of Missouri Interstate highway crash data have been utilized in this paper. The crash data comprises of traffic characteristics and geometric properties of highway segments.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Crash analysis research in general has focused on the estimation of traditional crash prediction models such as negative binomial (NB) and Poisson regression models and their generalized forms due to their relatively good fit to the crash

* Corresponding author. Tel.: +1 573 308 7845.
E-mail addresses: ma998@mail.mst.edu (M.A. Mohammadi), vsam@mst.edu (V.A. Samaranayake), ghbham@uaa.alaska.edu (G.H. Bham).
[1] Tel.: +1 573 341 4658.
[2] Tel.: +1 907 786 6053; fax: +1 907 786 1079.

http://dx.doi.org/10.1016/j.amar.2014.07.001
2213-6657/© 2014 Elsevier Ltd. All rights reserved.

data (Shankar et al., 1995; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000; Savolainen and Tarko, 2005; Mohammadi et al., 2014). Such crash prediction models take into account the crash frequency of a transportation facility (unit of analysis), such as an intersection or highway segment as a function of traffic flow and other crash-related factors. In these predictions, a greater amount of crash data, i.e. years of data, adds up to the reliability of the model estimates by reducing the standard errors (Lord and Persaud, 2000). However, the same unit generates multiple observations over the years that might be correlated due to unobserved effects related to specific entities that remain constant over time (Park and Lord, 2009; Castro et al., 2012; Bhat et al., 2014; Mannering and Bhat, 2014; Zou et al., 2014). In fact, these unobserved effects create a serial correlation in the repeated observations from the same unit over the years. Serial correlation in longitudinal data is an important issue, as it violates the independence assumptions on unobserved error terms in Poisson and/or NB crash frequency models, and creates inefficiency in the coefficient estimations and bias (underestimation) in estimation of standard error (Ulfarsson and Shankar, 2003; Washington et al., 2011; Dupont et al., 2013; Mohammadi et al., 2013; Bhat et al., 2014; Xiong et al., 2014).

Marginal models appear to be the most appropriate models for handling the temporal correlation, such as the work of Maher and Summersgill (1996) that uses an iterative solution based on the method of "constructed variables" presented by McCullagh and Nelder (1989). However, the extent and type of temporal correlation requires prior information that is not always known to the analyst (Lord and Persaud, 2000). Ulfarsson and Shankar (2003) tried to address the unit-specific serial correlation issue by using negative multinomial (NM) models in panel data and comparing the results with NB and random-effect negative binomial (RENB) model estimates. They showed that when there is correlation in the segment specific observations, the NM model is a much better fit compared to NB and RENB models. Dong et al. (2014a,2014b,2014c) developed multivariate random parameter models to account for the correlated crash frequency data as a result of unobserved heterogeneity. However, the model estimation methodology – the full Bayesian method – is complex, and the implementation and transferability of the method is not straightforward. Other research studies have been conducted in road safety analysis to account for such correlations in longitudinal data, yet consequences of the omission of the serial correlation are still not completely known. The most recent studies using longitudinal crash data include the work conducted by Venkataraman et al. (2014) to develop random parameter negative binomial models to investigate heterogeneity in crash means and the effects of interchange type on crash frequency.

Negative binomial models with a trend variable have also been used to study crash data with temporal correlation (Lord and Persaud, 2000; Noland et al., 2008; Quddus, 2008; Chi et al., 2012). Wang and Abdel-Aty (2006) used the technique of generalized estimating equations (GEE) to model rear-end crash frequencies at signalized intersections in order to account for the temporal and/or spatial correlation. GEE treats each highway segment as a cluster whose crash frequency observations have a temporal correlation over multiple years. In statistical terms, GEE captures the correlation incorporated in the error terms for model estimation. Hanley et al. (2003) showed that the use of GEE has the advantage of producing reasonably accurate standard errors and confidence intervals, especially when there are many subjects and few events. Hutchings et al. (2003) compared the performance of GEE with logistic regression by examining the change in parameter and variance estimates and the statistical significance of the independent variables. They found a lower number of significant variables when using the GEE method, and so recommended the use of nested structure models and GEE for analyzing motor vehicle crashes. Chang et al. (2006) applied the GEE procedure in a study of the effectiveness of drivers' license revocation and its impact on offenders in Taiwan. Lenguerrand et al. (2006) used multilevel logistic models (MLM), GEE, and logistic models (LM) to analyze hierarchical correlated crash data structure and found that both GEE and LM underestimate the parameters and confidence intervals, making MLM the most efficient model followed by GEE and LM models.

Lord and Mahlawat (2009) used GEE method with an autoregressive (AR) correlation structure to investigate the effect of a small sample size and low mean value of crash frequency on the reliability of the inverse dispersion parameter estimate. They found that the standard errors of the models' coefficients are larger when the serial correlation is accounted for in the modeling process. Méndez et al. (2010) used both logistic regression and GEE models (with exchangeable correlation structure) to study the relationship of a car's registration year and its crashworthiness. Peng et al. (2012) also utilized the GEE method with an exchangeable correlation structure to study the relationship between drivers' inattention and their inability in lane keeping. Stavrinos et al. (2013) used a GEE Poisson regression to study the impact of various distractions on driving behavior. Since the GEE models are not based on maximum likelihood estimation (MLE), they used a Chi-square test to estimate the significance of the variables. Giuffrè et al. (2013) studied the concepts of dispersion and correlation in yearly crash frequency data and presented a quasi-Poisson model in a GEE framework to incorporate both the dispersion and temporal correlation. In comparing the GEE with the COM-Poisson regression model, they recommended the use of GEE whenever it is handy. GEE procedure is robust against misspecification of the correlation structure in the response variable, but in that case, one may lose significant model efficiency and cause a misleading interpretation of the results, which in turn affects the reliability of the final safety estimation (Giuffrè et al., 2013).

The examples outlined above illustrate how GEE is not actually a regression model, but rather a method used to estimate models for data characterized by serial correlation. Throughout this paper, the models with temporal correlation that use GEE procedure are referred to as the GEE models. Unlike the traditional marginal models, the GEE models can handle temporal or other forms of correlation, even if the extent and type of correlation is unknown. However, Giuffrè et al. (2007) demonstrated that utilizing data correlation structure in safety modeling results in higher estimation precision. Although they have acknowledged that GEE models generally are robust to misspecification of the

Download English Version:

https://daneshyari.com/en/article/1104514

Download Persian Version:

https://daneshyari.com/article/1104514

Daneshyari.com