Analytic Methods in Accident Research 1 (2014) 39–52

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models

Yajie Zou [1], Yunlong Zhang [*,2], Dominique Lord [3]

Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, USA

ARTICLE INFO

Article history:
Received 27 June 2013
Received in revised form
1 November 2013
Accepted 1 November 2013
Available online 9 December 2013

Keywords:
Negative binomial
Finite mixture model
Dispersion
Weight parameter
Crash data
Heterogeneity

ABSTRACT

Factors that cause heterogeneity found in motor vehicle crash data are often unknown to transportation safety researchers and failure to capture this heterogeneity in statistical models can weaken the validity of modeling results. A finite mixture of regression models has been proposed to address the unobserved heterogeneity in crash data, and a fixed weight parameter for these models (i.e., the weight parameter is invariant of the characteristics of the observations under study) is commonly assumed. Recent studies have found that the weight parameter of the finite mixture of negative binomial (NB) models can be dependent upon the functional form of the attributes of the sites, and the selection of the functional form for weight parameter has a significant impact on the classification results.

This study investigates the effect of different functional forms on the estimation of the weight parameter as well as the group classification of the finite mixture of NB regression models, using crash data collected on rural roadway sections in Indiana. A total of 11 different functional forms for the varying weight parameter were estimated; these functional forms include various combinations of traffic flow and segment length as covariates. The results suggest that the modeling of the weight parameter (which essentially helps in improving the group classification) is generally necessary when using the finite mixture of NB regression models to analyze the crash data, even in the presence of a well-defined mean function. This study also confirms that the selection of the functional form for weight parameter will affect the classification results significantly. Among 11 different functional forms, one functional form, which uses the linear combination of different explanatory variables to model the classification, stands out based on both the goodness-of-fit statistics and the classification results, and is recommended for describing the weight parameter when using the finite mixture of NB regression models with varying weight parameters to analyze crash data.

1. Introduction

Research on understanding the factors that affect the probability of vehicle crashes has been of great interest to transportation safety analysts for many years. A number of factors known to influence traffic safety are driving-related

* Corresponding author.
E-mail addresses: zouyajie@gmail.com (Y. Zou), yzhang@civil.tamu.edu (Y. Zhang), d-lord@tamu.edu (D. Lord).
[1] Tel.: +1 936 245 5628; fax: +1 979 845 6481.
[2] Tel.: +1 979 845 9902; fax: +1 979 845 6481.
[3] Tel.: +1 979 458 3949; fax: +1 979 845 6481.

2213-6657/$ - see front matter
http://dx.doi.org/10.1016/j.amar.2013.11.001

factors (acceleration, braking and steering information, driver response to unexpected incidents and so on); factors related to road and vehicle (roadway geometric configurations, surface conditions, vehicle features, etc.); traffic and environment-related factors (traffic flows and composition, traffic control, weather conditions, etc.). Unfortunately, some factors (i.e., driving-related data) are usually not observable. As a result, researchers mainly focus on investigating the observed factors that affect the number of crashes for roadway segments or intersections over some fixed time periods (Lord and Mannering, 2010). For the crash count data, to overcome the over-dispersion problem associated with the Poisson regression, various ways have been proposed within the negative binomial (NB) models (Poch and Mannering, 1996; Hauer, 2001; Miaou and Lord, 2003; El-Basyouny and Sayed, 2006; Mitra and Washington, 2007; Malyshkina et al. 2009; Anastasopoulos and Mannering, 2009; El-Basyouny and Sayed, 2010). Recently, a quantile regression method was introduced by Qin and Reyes (2011) to analyze the heterogeneous crash data. The quantile regression can offer a complete view of how the explanatory variables affect the crash occurrence from the full range of the distribution.

Since the occurrence of crashes is rare (relatively speaking), to ensure the adequacy of sample size for valid and robust statistical estimation and inferences, crash data are often aggregated from a wide range of geographic locations. The aggregated crash data may contain heterogeneity. Heterogeneity implies that crash data may be collected from different sources (i.e., crash data collected at similar locations may share some common characteristics, while crash data collected at different locations may exhibit different characteristics). Thus, it is reasonable to assume that the sites with different combinations of factors (i.e., geometric design features, etc.) can constitute distinct sub-populations (sites are heterogeneous across and homogeneous within the sub-populations). Under this assumption, the commonly used negative binomial (NB) regression model may become inappropriate and the model estimation and inference from an NB modeling framework could be inefficient or misleading. Therefore, as proposed by Park and Lord (2009), it is reasonable to hypothesize that the individual crashes on highway entities (intersections or road segments) are generated from a certain number of hidden subgroups that are unknown to transportation safety researchers.

The concept of finite mixture distribution can date back to 1943 (Frühwirth-Schnatter, 2006). So far, the mixture modeling techniques have been applied in some transportation research areas. For example, to capture the heterogeneity in speed data, the normal mixture model has been used to fit the distribution of speed (Jun, 2010; Park et al., 2010b; Zou and Zhang, 2011; Zou et al., 2012). A finite mixture of regression models has been proposed to address the over-dispersion problem in crash data (Park and Lord, 2009). Xiong and Mannering (2013) developed a finite mixture random-parameters approach to study the heterogeneous effects of guardian supervision on crash-injury severities. For a standard finite mixture of regression models, previous studies (Park and Lord, 2009; Park et al., 2010a; Chang and Kim, 2012) have used a fixed weight parameter that is applied to the entire dataset. Latent segmentation models, which are similar to the finite mixture model, have been developed and applied in various transportation studies (Bhat, 1997; Greene and Hensher, 2003; Eluru et al., 2012; Sobhani et al., in press). These studies used different exogenous variables to estimate the weight parameter in the latent segmentation models. Recently, the finite mixture of NB regression models with varying weight parameters has been introduced by Zou et al. (2013) for analyzing the dispersed crash data. In their study, the weight parameter of the finite mixture models is assumed to be variable and can be dependent upon the attributes of the sites (i.e., covariates), such as segment length, traffic flow, etc. The results suggest that the two-component finite mixture of NB regression models (termed as the FMNB-2 model) with varying weight parameters can provide more reasonable classification results, as well as better statistical fitting performance than the FMNB-2 models with fixed weight parameters. Zou et al. (2013) also noted that the selection of the functional form for weight parameter has a significant impact on the classification results. For the FMNB-2 models with varying weight parameters, suspicious modeling results and erroneous inferences can be obtained if the functional form for weight parameter is mis-specified. Thus, there is a need to investigate how different functional forms affect the estimation of the varying weight parameter and whether there is a common functional form that can be properly used to model the weight parameter for different crash datasets. So far, a few functional forms have been proposed for estimating the weight parameter that varies across observations.

The primary objective of this research is to investigate the effect of different functional forms on estimation of the weight parameter as well as the group classification. Specifically, we mainly examine the modeling results and group classification from the finite mixture of NB regression models with different functional forms for the varying weight parameter. To accomplish the study objectives, 11 different functional forms for the varying weight parameter are estimated using the crash data collected on rural road sections in Indiana.

2. Background

This section describes the characteristics of the finite mixture of NB regression models with fixed and varying weight parameters.

2.1. Finite mixture of NB regression models

This study adopts a finite mixture model to describe heterogeneous crash data. To deal with the unobserved heterogeneity, the occurrence of the random vector, $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ is assumed to follow a finite mixture distribution. The mixture model is very flexible and the probability density function of a g-component mixture distribution can be