Latent segmentation based count models: Analysis of bicycle safety in Montreal and Toronto

Shamsunnahar Yasmin, Naveen Eluru *

Department of Civil, Environmental & Construction Engineering, University of Central Florida, United States

ABSTRACT

The study contributes to literature on bicycle safety by building on the traditional count regression models to investigate factors affecting bicycle crashes at the Traffic Analysis Zone (TAZ) level. TAZ is a traffic related geographic entity which is most frequently used as spatial unit for macroscopic crash risk analysis. In conventional count models, the impact of exogenous factors is restricted to be the same across the entire region. However, it is possible that the influence of exogenous factors might vary across different TAZs. To accommodate for the potential variation in the impact of exogenous factors we formulate latent segmentation based count models. Specifically, we formulate and estimate latent segmentation based Poisson (LP) and latent segmentation based Negative Binomial (LNB) models to study bicycle crash counts. In our latent segmentation approach, we allow for more than two segments and also consider a large set of variables in segmentation and segment specific models. The formulated models are estimated using bicycle-motor vehicle crash data from the Island of Montreal and City of Toronto for the years 2006 through 2010. The TAZ level variables considered in our analysis include accessibility measures, exposure measures, sociodemographic characteristics, socioeconomic characteristics, road network characteristics and built environment. A policy analysis is also conducted to illustrate the applicability of the proposed model for planning purposes. This macro-level research would assist decision makers, transportation officials and community planners to make informed decisions to proactively improve bicycle safety – a prerequisite to promoting a culture of active transportation.

© 2016 Elsevier Ltd. All rights reserved.

1. Background

Active forms of transportation such as walking and bicycling have the lowest carbon footprint on the environment and improve the physical health of pedestrians and bicyclists. With growing concern of worsening global climate change and increasing obesity among adults in developed countries, it is hardly surprising that transportation decision makers are proactively encouraging the adoption of active forms of transportation for short distance trips. For instance, bicycling, as a transport mode, is experiencing increased patronage and support in most Canadian cities where personal automobiles are the most common mode of transportation. In fact, between 1996 and 2006, a 42% increase in the number of daily bike commuters was observed in Canada (Pucher et al., 2011).

However, transportation safety concerns related to active transportation users form one of the biggest impediments to their adoption as a preferred alternative to private vehicle use for shorter trips. Earlier research reveals that the likelihood of being involved in a collision increases as the number of cyclists on the road increases (Wei and Lovegrove, 2013). Also, the risk of being injured in a collision while cycling could be about seven times higher than a motorist (Reynolds et al., 2009). Thus, traffic crashes and the consequent injury and fatality remain a detriment for cycling, leading to low bicycle mode share, specifically in North American communities (Wei and Lovegrove, 2013). Any effort to reduce the social burden of these crashes and encourage people to use bicycle for their daily short trips would necessitate the implementation of policies that enhance safety for bicyclists. An important tool to identify the critical factors affecting occurrence of bicycle crashes is the application of planning level crash prediction models.

1.1. Earlier research

Traffic crashes aggregated at a certain planning scale, for any given time interval, are non-negative integer valued events.

* Corresponding author.
  E-mail addresses: Shamsunnahar.Yasmin@ucf.edu (S. Yasmin),
naveen.eluru@ucf.edu (N. Eluru).

Naturally, these integer counts are examined employing count regression approaches. The traditional Poisson regression and Negative Binomial (NB) models have been the workhorses in examining the crash count events in safety literature. A number of research efforts have examined transportation (vehicle, pedestrian and bicyclist) related crash frequency (see Lord and Mannering (2010) for a detailed review). These studies have been conducted for different modes – vehicle (automobiles and motorbikes), pedestrian and bicycle and for different scales – micro (such as intersection and segment) and macro-level (such as traffic analysis zone, county, census tract). It is beyond the scope of the paper to review all the research on transportation crash frequency (for example see Brüde and Larsson, 1993; Turner et al., 2006; Loo and Tsui, 2010; Carter and Council, 2007; Jung et al., 2014; Dong et al., 2014 for micro-level studies). In our paper, we focus on studies examining crashes at the planning/macro-level.

A summary of earlier studies investigating crash frequency at a macro-level is presented in Table 1. The information provided in the table includes the methodological approach employed, the spatial aggregation level considered and the variable categories considered in the analysis from six variable categories – accessibility measures, exposure measures, sociodemographic characteristics, socioeconomic characteristics, road network characteristics and built environment. The following observations can be made from the table. First, the most prevalent spatial unit considered at the macro-level analysis is Traffic Analysis Zone (TAZ).[1] Second, NB model is the most frequently used statistical technique for examining crashes at the aggregate level. Third, very few studies (5 out of 33) explored bicycle crash frequency at the planning level. Fourth, none of the studies have employed latent segmentation based approach in examining crash frequency at macro-level.

With respect to macro-level bicycle crash frequency, the overall findings from earlier research efforts are usually consistent. The most commonly identified variables that contribute to the increase in bicycle crash risk include: (1) accessibility measures such as transit accessibility and number of bus stops, (2) exposure measures such as households with no cars, population density and total bicycle commuters, (3) sociodemographic characteristics such as proportion of young population and African population, (4) socioeconomic characteristics such as low-income population and per capita expenditure on alcohol, (5) road network characteristics such as street connectivity, total number of intersections and total on-street bike lanes and (6) built environment characteristics such as neighborhood compactness, higher mix of land use and proximity to academic buildings.

1.2. Current study in context

The overview of earlier literature indicates that, in recent years, examining crash frequency at the macro-level has seen a revival of interest among safety researchers. However, there is paucity of research focusing on macro-level bicycle crashes. Therefore, it is important to investigate zonal level bicycle crashes to identify critical factors and propose implications to facilitate proactive safety-conscious planning. A critical component in the process of identifying the contributing risk factors is the application of appropriate econometric models. As indicated in Table 1, the most

prevalent formulation to study macroscopic crash frequency is the NB model. NB model allows for overdispersion and thus provides a natural enhancement over the traditional Poisson model and is easy to estimate with a closed form structure to accommodate for unobserved heterogeneity. However, NB model (and Poisson model) typically restricts the impact of exogenous variables to be same across the entire population of crash events – population homogeneity assumption. But, the impact of control variables on bicycle crash frequency might vary across TAZs based on different attributes. Ignoring such heterogeneous impact of variables might result in incorrect coefficient estimates.

To account for systematic heterogeneity, researchers have employed a clustering technique (Karlaftis and Tarko, 1998). In this approach, target groups are divided in to different clusters based on a multivariate set of factors and separate models are estimated for each cluster. However, the approach requires allocating data records exclusively to a particular cluster, and does not consider the possible effects of unobserved factors that may moderate the impact of observed exogenous variables. Additionally, this approach might result in very few records in some clusters resulting in loss of estimation efficiency. An alternative approach to accommodate for population heterogeneity is to employ random parameter count models (Ukkusuri et al., 2011). However, in this approach the focus is on incorporating unobserved heterogeneity through the error term which necessitates extensive amount of simulation for model estimation while also not considering for observed heterogeneity.

A possible work around to accommodate for population heterogeneity is the application of latent segmentation based approach (or sometimes also referred to as finite mixture model). In this approach TAZs are allocated probabilistically to different segments and a segment specific model is estimated for each segment. Such an endogenous segmentation scheme is appealing for many reasons: First, each segment is allowed to be identified with a multivariate set of exogenous variables, while also limiting the total number of segments to a number that is much lower than what would be implied by a full combinatorial scheme of the multivariate set of exogenous variables. Second, the probabilistic assignment to segments explicitly acknowledges the role played by unobserved factors in moderating the impact of observed exogenous variables. Finally, this approach is semiparametric and hence, there is no need to specify a distributional assumption for the coefficients as is required in random parameter models (Greene and Hensher, 2003; Yasmin et al., 2014).

To be sure, the latent segmentation approach has been employed recently in safety literature for examining traffic crash count events at micro-level (Park et al., 2010; Park and Lord, 2009; Zou et al., 2014). However, the role of such population heterogeneity, in the context of macro-level crash count models has not been investigated in the existing literature. The microscopic models were developed with either a fixed weight parameter or segmentation model with a very small number of parameters (in the segmentation and segment specific models) were estimated citing model estimation complexity challenges. Further, earlier micro-level studies restricted the number of segments to two without any model selection exercise. The current study enhances the methodology from earlier finite mixture based count models in two ways: (1) we consider a large set of exogenous variables in the segmentation and segment specific models, and (2) we estimate more than two segments of latent segmentation and provide a clear framework for model selection.[2]

---

[1] TAZ is a traffic related geographic entity delineated by state and/or local transportation officials. The geographic unit is used for recording traffic-related data (for instance journey-to-work, place of work records) and are most widely used for conventional transportation planning. The size of a zone varies and the layout of the zonal system is usually defined based on the similarities in socioeconomic and/or land use conditions. TAZ is the most frequently used as spatial unit for macroscopic crash risk analysis (Abdel-Aty et al., 2013), and hence TAZ is considered as the aggregate unit of analysis in the current study context.

[2] The framework proposed in our analysis has been employed for discrete outcome – ordered or unordered – analysis earlier (see Bhat, 1997; Eluru et al., 2012; Yasmin et al., 2014).