Accident Analysis and Prevention 67 (2014) 129–136

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Examining the nonparametric effect of drivers' age in rear-end accidents through an additive logistic regression model

Lu Ma [1], Xuedong Yan *

MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, PR China

ARTICLE INFO

Article history:
Received 26 September 2013
Received in revised form 27 February 2014
Accepted 28 February 2014
Available online 15 March 2014

Keywords:
Additive logistic regression
Relative risks
Age effect
Inherently matched pairs

ABSTRACT

This study seeks to inspect the nonparametric characteristics connecting the age of the driver to the relative risk of being an at-fault vehicle, in order to discover a more precise and smooth pattern of age impact, which has commonly been neglected in past studies. Records of drivers in two-vehicle rear-end collisions are selected from the general estimates system (GES) 2011 dataset. These extracted observations in fact constitute inherently matched driver pairs under certain matching variables including weather conditions, pavement conditions and road geometry design characteristics that are shared by pairs of drivers in rear-end accidents. The introduced data structure is able to guarantee that the variance of the response variable will not depend on the matching variables and hence provides a high power of statistical modeling. The estimation results exhibit a smooth cubic spline function for examining the nonlinear relationship between the age of the driver and the log odds of being at fault in a rear-end accident. The results are presented with respect to the main effect of age, the interaction effect between age and sex, and the effects of age under different scenarios of pre-crash actions by the leading vehicle. Compared to the conventional specification in which age is categorized into several predefined groups, the proposed method is more flexible and able to produce quantitatively explicit results. First, it confirms the U-shaped pattern of the age effect, and further shows that the risks of young and old drivers change rapidly with age. Second, the interaction effects between age and sex show that female and male drivers behave differently in rear-end accidents. Third, it is found that the pattern of age impact varies according to the type of pre-crash actions exhibited by the leading vehicle.

1. Introduction

Age has been identified as a crucial factor that affects the behavior of driving and is consequently associated with the risks of causing or being involved in accidents. Earlier studies (e.g. Massie et al., 1995; Zhang et al., 1998) observed that young and old drivers are more likely to be involved in crashes and that old drivers also present the highest fatal involvement rate. Based on regression analyses, many other studies (e.g. Abdel-Aty and Radwan, 2000; Yan et al., 2005) have discovered similar patterns by examining the risk-taking behaviors of drivers in different age groups. Generally, the overall age effect on accident risk exhibits a U-shaped trend. The underlying reasons are, firstly, that old drivers have slower perception and reaction times (Abdel-Aty et al., 1998), which may lead to an excessive load from mental activities while driving (Cantin et al., 2009). On the other hand, young drivers are more likely to exhibit less maturity and less experience in their driving skills (Borowsky et al., 2010), as well as speeding behaviors.

Past accident analyses have largely relied on parametric statistical models, applied to observed crash data from police reports or experimental data from driving simulators (e.g. Martin et al., 2010) or eye tracers (e.g. Borowsky et al., 2010). In these models, the nonlinear effect of age is formulated by categorizing age into several predefined groups (e.g. Yan et al., 2005; Martin and Lenguerrand, 2008) or using a quadratic spline function with predefined knots (Cummings et al., 2003b). In addition to these considerations, many other aspects have been applied to enrich the inspection of the age effect. For example, Lourens et al. (1999) performed a multivariate analysis with correction for annual mileage, and Kim et al. (2013) took the heterogeneous effects of age into account. Gwyther and Holland (2012) introduced a theory of the self-regulation behavior of age, in which old drivers tend to avoid driving in poor situations due to compromised physical condition.

Undeniably, past studies have explored the age effect with various considerations. However, a more natural and flexible way of formulating the nonlinear effect of age is still lacking because

* Corresponding author. Tel.: +86 1051684602; fax: +86 1051840080.
E-mail addresses: lma@bjtu.edu.cn (L. Ma), xdyan@bjtu.edu.cn (X. Yan).
[1] Tel.: +86 1051684599; fax: +86 1051840080.

http://dx.doi.org/10.1016/j.aap.2014.02.021
0001-4575/

inappropriate specifications can introduce additional bias into the estimation. For example, the results from models in which age is categorized into predefined groups will depend on the arbitrary definition of these groups. In addition, recent findings have indicated that the proportions of licensed drivers within different age groups have changed noticeably in many countries over the years (Sivak and Schoettle, 2012) and more notably the percentage of licensed drivers in the older group has increased rapidly for the last several decades. Therefore, it is necessary to inspect the mechanism used to connect accident risk to age through more flexible and precise specifications.

Treated as a continuous variable, age has been examined with nonparametric specifications in many medical and epidemiological studies for a long time. For example, Durrleman and Simon (1989) characterized the nonlinear relationship between age and survival using Stanford Heart Transplant data, and Hultman et al. (2011) used smoothing splines to model the association between paternal age and offspring autism. However the nonparametric nature of age has not been investigated in traffic accident analyses. To this end, this study seeks to introduce an alternative approach based on a flexible specification of age in a more objective manner.

An immediate impediment to developing models to examine age impact would be that the observed data do not cover those drivers who are not involved in any crashes during the research period, because crash data usually originate from police reports of accidents (Lourens et al., 1999). In this study, an inherently matched paired data structure is developed to overcome this issue, by matching the driver who was not at fault (struck) with the driver who was at fault (striking) in the same accident. The method is effective for evaluating relative accident risks by examining the characteristics of drivers. Similar to the case–control analysis (Agresti, 2002) used in epidemiology studies, the observations in a matched pair here consist of exactly one driver whose role is being at fault (striking) and one driver whose role is not being at fault (struck). The method actually also belongs to the quasi-induced conception (Yan et al., 2005). Therefore, this study is restricted to rear-end accidents involving only two vehicles, partly because such data are perfect for constituting the paired samples.

Additive models are designed to supply nonparametric specifications for measuring the nonlinear effects of factors through flexible additive terms. In order to formulate a regression model able to capture more natural patterns of age impact, smooth functions of age are adopted as an additive term in this study, replacing the traditional linear term under fixed coefficients. Under such a modification, the additive models essentially constitute a more general family of models, which is advantageous in considering age impact.

Generalized additive models (GAM) (Hastie and Tibshirani, 1986) provide an extension to the generalized linear models, including such additive terms. Under the GAM framework, past studies have attempted to model crash frequency using negative binomial additive models (Li et al., 2011; Xie and Zhang, 2008). Yet, these models were based on subjects at more aggregated levels, for example accident counts of certain transportation facilities, and it is difficult to use these models to measure the age impact related to risk-taking behaviors for disaggregated subjects (drivers). Consequently, this study introduces the nonparametric version of regression models for examining drivers' rear-end crash-involvement risk using an inherently matched paired data structure. On the other hand, age itself needs to be treated as a continuous variable in order for the sensitivity between aging and the exhibited risks of drivers to be understood, which has usually been ignored by past studies (e.g. Yan et al., 2005).

2. Method

2.1. Additive logistic regression

The roles of drivers in rear-end accidents are treated as a dichotomous response variable in the following analysis. Thus, conditional on the occurrence of rear-end accidents, the outcome for a particular driver may be affected by his/her age as well as other attributes. A logistic regression structure with an additive term on age is therefore suggested.

The traditional logistic regression is a member of the set of generalized linear models (McCullagh and Nelder, 1989) with the link function in a logit form. The log odds for observing an outcome $Y=1$ are assumed to be a linear combination of all explanatory variables. The link function of a traditional logistic regression is presented in Eq. (1), and with such a specification, the magnitudes of the coefficients $\boldsymbol{\beta}$ can be used to reflect the association between the response variable Y and the explanatory variables $\boldsymbol{x}$. Particularly for a dummy variable x_i, the value $\exp(\beta_i)$ is exactly the odds ratio between the two groups of populations in the two statuses of x_i. If x_i is a continuous variable, $\exp(\beta_i)$ is the odds ratio between any two groups of populations if they differ by exactly one unit in x_i irrespective of the absolute position of the value of the variable. For the logistic regression model, these coefficient parameters can be estimated by maximizing the likelihood function.

$$\log\left[\frac{P(Y=1|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{1}$$

A general extension of traditional logistic regression replaces the linear terms with more flexible functions of the continuous explanatory variables. It is also called the additive logistic regression model (Hastie et al., 2009) and the revised link function is presented in Eq. (2). As aforementioned, such an extension greatly enhances the flexibility of the pattern of influence from the continuous explanatory variables. Meanwhile, the degrees of freedom of the model can be large, due to the design of these functions, which could lead to the issue of over-fitting on the training data.

$$\log\left[\frac{P(Y=1|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right] = \beta_0 + f_1(x_1) + \cdots + f_p(x_p). \tag{2}$$

The traditional MLE (maximum likelihood estimation) approach is not applicable to this analysis due to the excessive number of parameters, and instead Eq. (3) provides a general form of penalized likelihood function designed to overcome this estimation issue. $L(\boldsymbol{f}; \boldsymbol{\lambda})$ is the penalized likelihood function for additive logistic regression models, and consists of two parts. The first part is the ordinary log-likelihood function, which measures the closeness between the data and the fitted models. In order to avoid over-fitting, the second part penalizes the overall curvature of the smooth function for each knot. $\lambda_j \in [0, \infty)$ is a fixed smoothing parameter used to balance the data fitting and the smoothness of the function $f_j(\cdot)$. For extreme situations, $\lambda_j = 0$ indicates the closest match between the data and the fitted model and, if $\lambda_j = \infty$, $f_j(\cdot)$ is linear in the jth explanatory variable since its second derivatives are forced to be zero at all points in order to achieve the maxima of $L(\boldsymbol{f}; \boldsymbol{\lambda})$.

$$\begin{aligned} L(\boldsymbol{f};\boldsymbol{\lambda}) &= \sum_{i=1}^{n} \{y_i \log P(Y=1|\boldsymbol{x}_i) + (1-y_i)\log[1-P(Y=1|\boldsymbol{x}_i)]\} \\ &\quad - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \int [f_j''(t)]^2 dt \\ &= \sum_{i=1}^{n} \{y_i[\beta_0 + \sum_{j=1}^{p} f_p(x_{ij})] - \log[1 + \exp(\beta_0 \\ &\quad + \sum_{j=1}^{p} f_p(x_{ij}))]\} - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \int [f_j''(t)]^2 dt \end{aligned} \tag{3}$$