PERGAMON

Accident Analysis and Prevention 34 (2002) 729–741

ACCIDENT
ANALYSIS
&
PREVENTION

www.elsevier.com/locate/aap

Using logistic regression to estimate the influence of accident factors on accident severity

Ali S. Al-Ghamdi *

College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia

Received 13 June 2000; received in revised form 28 May 2001; accepted 1 June 2001

Abstract

Logistic regression was applied to accident-related data collected from traffic police records in order to examine the contribution of several variables to accident severity. A total of 560 subjects involved in serious accidents were sampled. Accident severity (the dependent variable) in this study is a dichotomous variable with two categories, fatal and non-fatal. Therefore, each of the subjects sampled was classified as being in either a fatal or non-fatal accident. Because of the binary nature of this dependent variable, a logistic regression approach was found suitable. Of nine independent variables obtained from police accident reports, two were found most significantly associated with accident severity, namely, location and cause of accident. A statistical interpretation is given of the model-developed estimates in terms of the odds ratio concept. The findings show that logistic regression as used in this research is a promising tool in providing meaningful interpretations that can be used for future safety improvements in Riyadh.

Keywords: Logistic regression; Accident severity

1. Introduction

Accident severity is of special concern to researchers in traffic safety since this research is aimed not only at prevention of accidents but also at reduction of their severity. One way to accomplish the latter is to identify the most probable factors that affect accident severity. This study aims at examining not all factors, but some believed to have a higher potential for serious injury or death, such as accident location, type, and time; collision type; and age and nationality of the driver at fault, his license status, and vehicle type. Other factors were not examined because of substantial limitations in the data obtained from accident reports. Logistic regression was used in this study to estimate the effect of the statistically significant factors on accident severity. Logistic regression and other related categorical-data regression methods have often been used to assess risk factors for various diseases. However, logistic regression has been used as well in transportation studies. A brief literature review follows of the use of this type of regression in traffic safety research.

2. Literature review

Regression methods have become an integral component of any data analysis concerned with the relationship between a response variable and one or more explanatory variables. The most common regression method is conventional regression analysis (CRA), either linear or nonlinear, when the response variable is continuous (iid). However, when the outcome (the response variable) is discrete, CRA is not appropriate. Among several reasons, the following two are the most significant:

1. The response variable in CRA must be continuous, and
2. The response variable in CRA can take nonnegative values.

These two primary assumptions are not satisfied when the response variable is categorical.

* Tel.: +966-1-467-7019; fax: +966-1-467-4254.
E-mail address: asghamdi@ksu.edu.sa (A.S. Al-Ghamdi).

0001-4575/02/$ - see front matter
PII: S0001-4575(01)00073-2

Jovanis and Chang (1986) found a number of problems with the use of linear regression in their study applying Poisson regression as a means to predict accidents. For example, they discovered that as vehicle-kilometers traveled increases, so does the variance of the accident frequency. Thus, this analysis violates the homoscedasticity assumption of linear regression.

In a well-summarized review of models predicting accident frequency, Milton and Mannering (1997) state: "The use of linear regression models is inappropriate for making probabilistic statements about the occurrences of vehicle accidents on the road." They showed that the negative binomial regression is a powerful predictive tool and one that should be increasingly applied in future accident frequency studies.

Kim et al. (1996) developed a logistic model and used it to explain the likelihood of motorists being at fault in collisions with cyclists. Covariates that increase the likelihood of motorist fault include motorist age, cyclist age (squared), cyclist alcohol use, cyclists making turning actions, and rural locations.

Kim et al. (1994) attempted to explain the relationship between types of crashes and injuries sustained in motor vehicle accidents. By using techniques of categorical data analysis and comprehensive data on crashes in Hawaii during 1990, a model was built to relate the type of crash (e.g. rollover, head-on, sideswipe, rear-end, etc.) to a KABCO injury scale. They also developed an 'odds multiplier' that enabled comparison according to crash type of the odds of particular levels of injury relative to noninjury. The effects of seatbelt use on injury level were also examined, and interactions among belt use, crash type, and injury level were considered. They discussed how log-linear analysis, logit modeling, and estimation of 'odds multipliers' may contribute to traffic safety research.

Kim et al. (1995) built a structural model relating driver characteristics and behavior to type of crash and injury severity. They explained that the structural model helps to clarify the role of driver characteristics and behavior in the causal sequence leading to more severe injuries. They estimated the effects of various factors in terms of odds multipliers — that is, how much does each factor increase or decrease the odds of more severe crash types and injuries.

Nassar et al. (1997) developed an integrated accident risk model (ARM) for policy decisions using risk factors affecting both accident occurrences on road sections and severity of injury to occupants involved in the accidents. Using negative binomial regression and a sequential binary logit formulation, they developed models that are practical and easy to use. Mercier et al. (1997) used logistic regression to determine whether either age or gender (or both) was a factor influencing severity of injuries suffered in head-on automobile collisions on rural highways.

Logistic regression was also used by Hilakivi et al. (1989) in predicting automobile accidents of young drivers. They examined the predictive values of the Cattel 16-factor personality test on the occurrence of automobile accidents among conscripts during 11-month military service in a transportation section of the Finnish Defense Forces.

James and Kim (1996) developed a logistic regression model to describe the use of child safety seats for children involved in crashes in Hawaii from 1986 through 1991. The model reveals that children riding in automobiles are less likely to be restrained, drivers who use seat belts are far more likely to restrain their children, and 1- and 2-year-olds are less likely to be restrained.

3. Theoretical background of logistic regression

It is important to understand that the goal of an analysis using logistic regression is the same as that of any model-building technique used in statistics: to find the best fit and the most parsimonious one. What distinguishes a logistic regression model from a linear regression model is the response variable. In the logistic regression model, the response variable is binary or dichotomous. The difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression analysis. In any regression analysis the key quantity is the mean value of the response variable given the values of the independent variable:

$$E(Y/x) = \beta_0 + \beta_1 x$$

where Y denotes the response variable, x denotes a value of the independent variable, and the β_i-values denote the model parameters. The quantity is called the conditional mean or the expected value of Y given the value of x. Many distribution functions have been proposed for use in the analysis of a dichotomous response variable (Hosmer and Lemeshow, 1989; Agresti, 1984; Feinberg, 1980). The specific form of the logistic regression model is

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{1}$$

where, to simplify the notation, $\pi(x) = E(Y/x)$. The transformation of the $\pi(x)$ logistic function is known as the logit transformation:

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x \tag{2}$$