Accident Analysis and Prevention 121 (2018) 157–165

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A discrete mixture regression for modeling the duration of non-hospitalization medical leave of motor accident victims

Lluís Bermúdez[a], Dimitris Karlis[b], Miguel Santolino[a,*]

[a] University of Barcelona, Riskcenter-IREA, Spain
[b] Athens University of Economics and Business, Department of Statistics, Greece

ARTICLE INFO

Keywords:
Motor accident
Multiple negative binomial
Multiple Poisson
Work disability days

ABSTRACT

Studies analyzing the temporary repercussions of motor vehicle accidents are scarcer than those analyzing permanent injuries or mortality. A regression model to evaluate the risk factors affecting the duration of temporary disability after injury in such an accident is constructed using a motor insurance dataset. The length of non-hospitalization medical leave, measured in days, following a motor accident is used here as a measure of the severity of temporary disability. The probability function of the number of days of sick leave presents spikes in multiples of five (working week), seven (calendar week) and thirty (month), etc. To account for this, a regression model based on finite mixtures of multiple discrete distributions is proposed to fit the data properly. The model provides a very good fit when the multiples for the working week, week, fortnight and month are taken into account. Victim characteristics of gender and age and accident characteristics of the road user type, vehicle class and the severity of permanent injuries were found to be significant when accounting for the duration of temporary disability.

1. Introduction

Road traffic accidents are a major health problem worldwide and the eighth leading cause of death (WHO, 2013). The risk factors associated with the mortality and permanent injuries resulting from such accidents have been widely investigated in the literature (Shibata and Fukuda, 1994; Savolainen et al., 2011; Boucher and Santolino, 2010; Mannering and Bhat, 2014; Alemany et al., 2013; Tay and Rifaat, 2007; Yasmin and Eluru, 2013). However, studies analyzing the temporary consequences of motor vehicle accidents are more scarce. The period that motor victims are recovering from injuries has an important socioeconomic impact in terms of the use of health services and lost of productivity, among other consequences (Miller and Galbraith, 1995; Blincoe et al., 2002). This paper proposes a regression model to evaluate the risk factors affecting the duration of temporary disability as a result of road traffic injuries.

Temporary disability can be defined as the impairment of an individual's mental or physical faculties that impede the victim from functioning normally for as long as they remain under treatment (or until their injuries have stabilized). The most common approach taken in the literature to analyze the severity of temporary disability is to consider the length of hospitalization (Gardner et al., 2007; Peek-Asa et al., 2011; Ayuso et al., 2015; Santolino et al., 2012; Guria, 1990), and to examine its relationship with the characteristics of the injury suffered and those of the victim.

Analyses of the duration of hospitalization are in part motivated by the availability of data. However, such an approach may underestimate the total social costs of a traffic injury. Non-serious injuries do not, as a rule, require hospitalization, but may nevertheless be associated with substantial temporary disability, the case, for example, of whiplash injuries (Buitenhuis et al., 2009). For this reason, Ebel et al. (2004) made simulated projections of the number of work days lost as a result of motor vehicle crashes and studied the factors that influenced a victim's return to work. Berecki-Gisolf et al. (2013), on the other hand, restricted their analysis of the work disability period to musculoskeletal and orthopedic traffic injuries.

In this study our attention is focused on the analysis of factors affecting the length of temporary disability without hospitalization. The length of hospitalization was excluded from the analysis, since key drivers of hospital length of stay have been already investigated (Ayuso et al., 2015; Santolino et al., 2012). A motor insurance claim dataset is used to evaluate the number of days of medical leave taken by accident victims. Medical leave is defined in this article as the out-of-hospital period taken by motor victims to recover from injuries or until their injuries have stabilized. In Spain, the period of non-hospitalized temporary disability as a consequence of a motor crash is set by doctors of

* Corresponding author at: Department of Econometrics, University of Barcelona, Diagonal 690, 08034 Barcelona, Spain.
E-mail address: msantolino@ub.edu (M. Santolino).

https://doi.org/10.1016/j.aap.2018.09.006
Received 29 January 2018; Received in revised form 15 August 2018; Accepted 7 September 2018

the public health system who determine the number of days of medical leave required by out-patients. This information is required by insurance companies to compute the motor insurance compensation. So, the period of medical leave refers to the out-of-hospital recovering time after the accident taken by any type of motor victim and not only by the working population.

The frequency distribution of the length of non-hospitalization temporary disability (measured in days) exhibits regular spikes at certain multiples. The periodic peaks observed in the frequency distribution could reflect the time scales used by doctors when determining the number of days of sick leave before the next scheduled medical examination. For example, a doctor is more likely to program a re-evaluation of the medical evolution of injuries in two weeks' time than in thirteen days. This decision may be because doctors think on a daily/weekly/monthly scale when scheduling patient evaluations, based on the severity of injuries and the number of days the patient will be off sick. The doctor's agenda constraints may also be a reason (i.e. the doctor only visits one day in the week). In fact, regularly spaced spikes in the frequency distribution are observed at multiples of 5, 7, 15 and 30.

Data with periodic peaks are observed in various applications. Examples include the misreporting of age (Siegel and Swanson, 2004; Camarda et al., 2008), number of cigarettes smoked (Wang et al., 2012) and duration of unemployment (Torelli and Trivellato, 1993; Wolff and Augustin, 2003). This phenomenon of rounding exact counts to even multiples of reported units is known as digit preference or heaping. The literature on this phenomenon assumes that data can be interpreted as indirect (or rounded) observations of a latent distribution. The goal usually pursued is to model the unobserved latent variable using smoothing methods (Camarda et al., 2008; Wang et al., 2012; Wang and Heitjan, 2008; Wang and Wertelecki, 2013).

A different modeling approach is proposed in this paper. We directly model the random variable with peaks rather than with an unobserved smoothed variable. The methodology for fitting frequency data with regular spikes is based on finite mixtures of discrete distributions of different multiplicities, as proposed by Bermúdez et al. (2017). This methodology is extended to the regression modeling analysis reported in this article. A discrete mixture regression model is developed to fit data with regular spikes conditioned on a set of covariates. The duration of temporary disability following a traffic accident is modeled, including, as explanatory variables, characteristics of the victim (gender and age) and the accident (road user type, vehicle class and severity of permanent injuries).

The article is organized as follows. The regression model is presented in the next section. Section 3 describes the data. The results are shown in Section 4. Concluding remarks are given in Section 5.

2. Regression model

2.1. Discrete distributions

Let $X \in \mathbb{N}$ be a discrete random variable that takes non-negative integer values including zero. In statistics, the most frequently used parametric distributions to model discrete random variables are the Poisson distribution and the negative binomial (NB) distribution (Boucher and Santolino, 2010). The probability function (pf) of the Poisson distribution with parameter λ, denoted as $P_1^p(\lambda)$, is given by

$$P_1^p(X=x)=\frac{\exp(-\lambda)\lambda^x}{x!},\qquad \lambda\geq 0,\quad x=0,1,2,\ldots$$

The Poisson distribution has the following moments,

$$E(X)=\lambda \quad\text{and}\quad \mathrm{Var}(X)=\lambda.$$

The Poisson distribution assumes variance equal to the mean and, hence, it has limitations when dealing with overdispersed data, i.e. when the sample variance exceeds the sample mean. In this context, the negative binomial distribution is often more adequate. The pf of the negative binomial distribution with parameter λ and r, where λ is the mean parameter and r the additional parameter to account for over-dispersion, is given by

$$P_1^{\mathrm{nb}}(X=x)=\left(\frac{r}{r+\lambda}\right)^r\frac{\Gamma(r+x)}{x!\,\Gamma(r)}\left(\frac{\lambda}{r+\lambda}\right)^x,\quad x=0,1,2,\ldots$$

The NB distribution has the following moments,

$$E(X)=\lambda \quad\text{and}\quad \mathrm{Var}(X)=\left(\lambda+\frac{\lambda^2}{r}\right),$$

It is easy to see that if $r\to\infty$ the negative binomial tends to the Poisson.

2.2. Multiple discrete distributions

Often, the variable of interest is the sum of lower-level units and we are specifically interested in analyzing the random variable measured in the lower level units. For example, a survey will ask how many packs of cigarettes the subject smokes per week, because this is easier to calculate than the actual number of cigarettes; however, the variable of interest in the study is the number of cigarettes (let's say twenty per pack). In this case, the variable of interest takes multiples of twenty (that is 0, 20, 40, …).

To deal with data measured on a different scale to the scale of interest, multiple discrete distributions are used. Such distributions are generalizations of the discrete distributions that allow for different multiplicities. The multiple discrete distribution versions of the Poisson and NB are introduced.

The pf of the multiple Poisson with multiplicity m and parameter λ, denoted as P_m^p, is as follows:

$$P_m^p(X=y)=\begin{cases}\frac{\exp(-\lambda)\lambda^x}{x!} & y=\mathrm{m}x,\quad x=0,1,2,\ldots\\ 0 & \text{otherwise.}\end{cases}$$

It is straightforward to obtain the first two moments:

$$E(X)=m\lambda \quad\text{and}\quad \mathrm{Var}(X)=m^2\lambda,$$

This generalization gives positive probability to points 0, m, $2m$, … and 0 elsewhere. So, the Poisson distribution can be understood as a particular case of the multiple Poisson distribution with a multiplicity equal to one, $m = 1$. Using a similar approach, we can define the multiple negative binomial distribution. The pf and first two moments of the multiple negative binomial with multiplicity m and parameters λ and r are as follows:

$$P_m^{\mathrm{nb}}(X=y)=\begin{cases}\left(\frac{r}{r+\lambda}\right)^r\frac{\Gamma(r+x)}{x!\,\Gamma(r)}\left(\frac{\lambda}{r+\lambda}\right)^x & y=\mathrm{m}x,\quad x=0,1,2,\ldots\\ 0 & \text{otherwise,}\end{cases}$$

$$E(X)=m\lambda \quad\text{and}\quad \mathrm{Var}(X)=m^2\cdot\left(\lambda+\frac{\lambda^2}{r}\right),$$

Note that the multiple Poisson distribution is also a limiting case of the multiple negative binomial distribution when $r\to\infty$.

2.3. A finite mixture discrete distribution

When the random variable of interest can be interpreted as resulting from different subpopulations/subgroups, the finite mixture distribution can be easily derived from distributions of the individual subpopulations/subgroups. Alternative mixtures of discrete distributions have been defined in the literature. For example, in the road safety literature, the well-known zero-inflated distribution is a mixture between a Bernoulli distributed random variable and a discretely distributed random variable, such as a Poisson or negative binomial distribution (Lord et al., 2005; Ayuso et al., 2015; Anastasopoulos, 2016;