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[c,*]

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

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.

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

***Corresponding Author.** Department of Econometrics, University of Barcelona, Diagonal 690, 08034-Barcelona, Spain. Tel.:+34-93-4020484; fax: +34-93-4021821; e-mail: msantolino@ub.edu

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 sociocconomic 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 severity of a temporary disability is closely associated with the length of the period during which the victim is on sick leave recovering from the injuries sustained in the accident. 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.

The aim of this study is to describe the distribution and determinants of temporary disability duration outcomes for any type of motor vehicle injury.

A motor insurance claim dataset is used to evaluate the number of days of medical leave taken by accident victims. 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). In this study our attention is focused on the analysis of factors affecting the length of temporary disability without hospitalization. In Spain, the period of non-hospitalized temporary disability as a consequence of a motor crash is set by doctors of the public health system who determine the number of days of medical leave required by out-patients. This information is required to compute the motor insurance compensation and, in case of the victim was also time off from work, the paid sick leave amount.

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 reevaluation 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 modelling 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 modelling 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 modelled, 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

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

The Poisson distribution has the following moments,

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

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 overdispersion, is given by

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

The NB distribution has the following moments,

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

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 $0, 20, 40, \ldots$).

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:

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

It is straightforward to obtain the first two moments:

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

This generalization gives positive probability to points $0, m, 2m, \ldots$ 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:

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

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

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; Boucher and Santolino, 2010; Shankar et al., 1997). Other, less frequently used, mixtures in the road safety literature include finite mixture distributions based on the combination of two or more discrete distributions (Park and Lord, 2009; Zou et al., 2014; Park et al., 2010, 2014).

Let's consider that the random variable of interest $Y \in \mathbb{N}$ is constructed as a mixture of pairwise independent discrete random variables $X_1, X_2, \ldots, X_K$, where X_j takes no negative integer values $\forall j$, and a categorical random variable Z that consists of K categories. That is,

$$Y = \sum_{j=1}^{K} \mathbb{1}_Z \cdot X_K$$

where $\mathbb{1}_Z$ takes a value of 1 when $Z = j$ and 0 otherwise, $j = 1, \ldots, K$. Denote by Y_i the count of some event for $i = 1, \ldots, n$ observations. If we assume that X_j is a Poisson distributed random variable with parameter λ_j for $j = 1, \ldots, K$, then the pf of Y is defined as,

$$P(Y_i = y) = P(y) = \sum_{j=1}^{k} \pi_j P_1^p(y|\lambda_j), \quad y = 0, 1, \ldots \tag{1}$$

where $\pi_j = P(Z_i = j)$ is the probability that the i-th observation belongs to group j with $0 < \pi_j < 1$ for $j = 1, \ldots, K$ with $\sum \pi_j = 1$. Intuitively, conditional on the fact that the i-th observation belongs to the j-th group or component, the observed counts come from a Poisson distribution with parameter λ_j. The probability that a randomly selected individual belongs to group j is $0 < \pi_j < 1$. The information provided by (1) is the unconditional probability when we do not know which group the i-th observation belongs to.

If we assume that X_j is a NB distributed random variable with parameters λ_j and r_j for $j = 1, \ldots, K$, then the pf of Y is defined as,

$$P(Y_i = y) = P(y) = \sum_{j=1}^{k} \pi_j P_1^{nb}(y|\lambda_j, r_j), \quad y = 0, 1, \ldots \tag{2}$$

As noted previously, an appealing feature of the multiple discrete distribution is that we can model data measured on different scales. Expressions (1) and (2) can be generalized for multiplicities different to one. If we assume that m_j is the multiplicity associated with the Poisson distributed X_j, then pdf of Y is defined as,

$$P(Y_i = y) = \sum_{j=1}^{K} \pi_j P_{m_j}^{p}(y|\lambda_j), \quad y = 0, 1, \ldots \tag{3}$$

If X_j is a NB distributed random variable, then

$$P(Y_i = y) = \sum_{j=1}^{K} \pi_j P_{m_j}^{nb}(y|\lambda_j, r_j), \quad y = 0, 1, \ldots \tag{4}$$

The idea underpinning this representation is that we can have data that are measured on different scales. For example, we know that the data are NB distributed, but we have observations measured in terms of days, weeks, and months (that is $P_1^{nb}(\lambda_1, r_1)$, $P_7^{nb}(\lambda_7, r_7)$ and $P_{30}^{nb}(\lambda_{30}, r_{30})$ respectively). So, if an observation is randomly selected, we have to consider the probability that the observation is measured on a particular scale (days, weeks, or months) and the probability that this observation takes a specific value given that it is measured on this particular scale. Obviously, each of the components has a simple interpretation when modelling and fitting the data referring to a particular scale (days, working weeks, weeks, and months). Here, it should be stressed that certain values of y might be obtained from different components. For example, in the case of the value $y = 35$, this might be obtained from components related to both day and week multiples, but not from components associated with the month multiple. Zero values can be seen in relation to all components.

The flexible construction of the mixture discrete distributions permits the combination of different families of discrete distributions, that is, some non-negative discrete random variables are Poisson distributed and others are

negative binomially distributed. An interesting issue emerges from definitions (3) and (4). In this article, we consider multiplicities as being known and, therefore, they can be set beforehand. Normally, it is possible to recognize the position of peaks in the frequencies, as this implies certain multiples related to the context and reflects our knowledge of the data. Bermúdez et al. (2017) show how multiplicities can be estimated when they are unknown.

2.4. A finite mixture discrete regression model

The goal is to construct a model to fit count data that exhibit periodic peaks in their frequency distribution attributable to the fact that different scales are being implicitly used at the same time. Models (3) and (4), or a combination of the two, can be extended if we allow the mean of each component to depend on some covariates related to the i-th individual.

We relate the λ's with some covariates. Hence we now have two subscripts for λ's assuming that

$$\log \lambda_{ij} = \beta_j' \mathbf{X}_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, K$$

where n is the sample size and β_j a vector of regression coefficients. The index j implies to which of the components they refer and $\mathbf{X}_{ij}$ is a vector of covariates for the i-th individual associated to the j-th component. For generality, we assume that a different vector of covariates is considered for each component. However, the modeling approach allows us to analyze the effect of the same covariates on the different time scales.

The mixing proportions may also be related to the covariate vectors by means of a standard multinomial logistic model. The interpretation of the mixing proportions is that they reveal the proportion of observations from each component and, hence, in this case, the time scale the doctors use when scheduling re-examinations.

2.5. Estimation

The estimation of the discrete mixture regression model is easily conducted using an expectation-maximization (EM) algorithm. In Bermúdez et al. (2017) an EM-type algorithm was described based on the fact that the model is represented as a standard finite mixture model. Here, we adapt the algorithm to the regression modeling context, but a closed form M-Step is no longer available. Instead, we fit a weighted GLM model at this stage.

Using the standard approach for finite mixtures, the set of unobserved latent component indicator variables Z_{ij} is defined, i.e., $Z_{ij} = 1$ if the i-th observation belongs to the j-th group and 0 otherwise. Hence $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ik})$. Note that some observations clearly have $Z_{ij} = 0$ by definition for certain multiplicities. For example, for the multiplicity equal to 2, every odd observation cannot be generated from this component.

The steps of the algorithm are the following.

At the E-step, using the current parameter estimates, and denoting as $P_{m_j}(\cdot|\theta_j)$ any model with multiplicity m_j and parameters θ_j, which can be either the Poisson or the negative binomial, calculate

$$w_{ij} = E(Z_{ij}) = \frac{\pi_j P_{m_j}(y_i|\theta_j)}{\sum_{r=1}^{K} \pi_r P_{m_r}(y_i|\theta_r)}$$

and then update the parameters at the M-step as

$$\hat{\pi}_j = \frac{\sum_{i=1}^{n} w_{ij}}{n}$$

which is always the case. For the regression and other parameters, this is achieved by fitting the relevant GLM model (Poisson or negative binomial) by weighted likelihood, using as weights the w_{ij}'s from the E-step, and using as responses the observed values divided by the multiplicity of the relevant components. Note that this will end up with an integer, because, if not, the weights w_{ij} will be 0.

As usual, the algorithm stops when the log-likelihood stops increasing. In practice, that is when the relative improvement is smaller than a small number (10^{-8} is used herein).

All the advantages and disadvantages of the EM for finite mixtures hold. But, of course, the typical cautionary remarks concerning EM also apply here. We need good initial values to avoid problems of local maxima. Simple initial values can be derived for considering the data from the possible multiplicities and from the initial values for λ's, as well as from the mixing proportions. For example, for a multiplicity $m = 10$, we can only use the data presenting values $10x$ and from these we can derive some estimate that fits a simple GLM. In our extensive examination of the algorithm, this approach leads to the rapid convergence of the algorithm to the global maximum.

Finally, note that the derivation of the finite mixture and the algorithm does not exclude the case of the same multiplicity for some components. In this case, the derivation of starting values is equivalent to that of finite Poisson mixtures.

3. Data

The data record the duration of outpatient medical leave owing to a motor vehicle accident. The database was provided by one of Spain's largest motor insurance companies. The data set has been used previously for other purposes in Boucher and Santolino (2010), Santolino et al. (2012) and Ayuso et al. (2016). We draw on 20,257 observations from non-fatal victims of traffic collisions in Spain. All of the victims included in the dataset received an insurance compensation for their injuries in 2007, although the accident may have occurred before that year. In the settlement year all of the individuals were fully recovered, or with stable injuries.

In Spain there exists a legislative scale regarding the assessment of damages for automobile bodily injuries in motor insurance claims. This scale provides a compensation system for three general categories: death, temporary disability and permanent disability. The severity of permanent injuries is assessed according to the motor victim's permanent disability score in accordance with the Spanish medical legal scale. The scale of permanent injuries ranges from 0 (no permanent injuries) to 100 points (maximum severity of permanent injuries). The judge determines the final injury score according to severity. The basic compensation for permanent injuries depends on the overall scoring (in positive proportion) and the age of the victim (in inverse proportion). Under the Spanish system the temporary disability entitle the victim to a daily basic compensation. The daily amount for temporary disability depends on whether the victim was hospitalized, out-of-hospital with inability to work or out-of-hospital without inability to work. The total basic compensation is obtained aggregating all compensatory components. The basic compensation is later adjusted according to the financial situation and family responsibilities of the victim. Information impacting the size of the insurance compensation is normally recorded by insurers when processing and tracking claims until settlement. Specifically, information regarding the number of days of disability reported by each victim is used to evaluate the amount of compensation for temporary disability. Unfortunately the standard information from the police accident reports could not be retrieved.

Police reports provide details of the circumstances of the collision but provide limited information on how injuries may develop after the collision.

The number of out-of-hospital disability days are shown in Figure 1. The most interesting feature of the data is that there occur various spikes at certain values. Figure 1 presents significant spikes at certain values that apparently coincide with the different time scales employed by doctors in evaluating the severity of the patient's injuries: 5 days (a working week), 7 days (a calendar week), 10 days, 15 days (a fortnight), 20 days, 30 days (a month), 60 days and 90 days (a trimester).

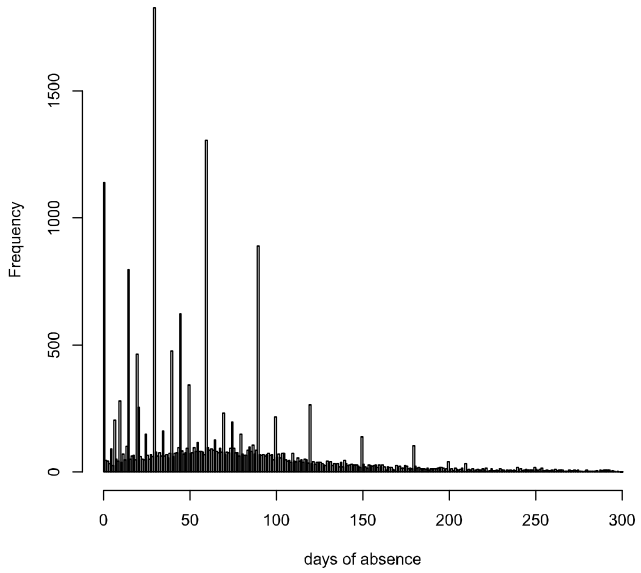

Figure 1: Empirical distribution of the number of days of temporary disability (limited to 300 days)

A set of regressors is considered when modelling this duration of temporary disability. A description of the variables is provided in Table 1. Explanatory factors include the victim's age and gender, casualty type (a distinction being drawn between drivers and non-drivers), vehicle type (a distinction between heavy and non-heavy vehicles) and the score (in log scale) for serious permanent injuries.

Gender	1 if the injured victim was male; 0 otherwise.
Age	Age of the victim in years (divided by 10).
Driver	1 if the injured victim was the driver; 0 otherwise.
Heavy vehicle	1 if the vehicle was a heavy vehicle; 0 otherwise.
Injury score	the log of the score (+1) for permanent injuries.

Table 1: Description of variables

	Mean	St.Dev	Min	Max
Days of temporary disability	77.78	82.70	0	995
Age	3.82	1.67	0	0.99
Injury score	1.20	0.72	0	4.61

(a)

	Variable equals 0 (Number of observations)	Variable equals 1 (Number of observations)
Gender	11,141	9,116
Driver	9,926	10,331
Heavy vehicle	19,285	972

(b)

Table 2: Descriptive statistics of continuous variables (a) and frequency distribution of binary variables (b)

Descriptive statistics of the continuous variables are recorded in Table 2a and the empirical frequency distribution of the dichotomous variables is shown in Table 2b. The empirical variance in the duration of temporary disability equals $6,839$ and the mean equals 77.78. That is, the sample variance is much greater than the sample expectation, indicating that the data show overdispersion.

4. Results

Model parameters were estimated via maximum likelihood (ML). The expectation-maximization algorithm described in section 2.5 was used to obtain ML estimates. Results were obtained in the programming language `R`.

4.1. Model selection

To start with, and in order to obtain a better initial idea of the modelling approach, several models without covariates, using increasingly more multiplicities, were fitted. In Table 3, the log-likelihood value and Bayesian Information Criterion (BIC) are reported for a set of discrete models without covariates combining different multiplicities. First, the Poisson and NB model with just the multiple associated with days ($m = 1$) were estimated, that is, the most common count data models. These models were labelled 1-Poisson and 1-NB, respectively, with the number indicating the multiplicity. In the next stage, the multiple associated with the working week ($m = 5$) was added to the 1-NB model and labelled 1-5 NB. In each new stage, an additional multiple was included. Weeks ($m = 7$), fortnights ($m = 15$), months ($m = 30$) and two months ($m = 60$) were considered.

Model	Log-lik	BIC
1-Poisson	-545226.60	1090463.12
1-NB	-106866.17	213752.17
1-5-NB	-101765.20	203579.98
1-5-7-NB	-100862.10	201803.53
1-5-7-15-NB	-98727.48	197564.04
1-5-7-15-30-NB	-97745.29	195629.41
1-5-7-15-30-60-NB	-97730.74	195630.06

Table 3: Comparison of the models (without covariates).

Table 3 shows that adding multiplicities notably improves the log-likelihood, indicating the need to account for the spikes in the data.

The next step involved adding the covariates to the regression model. The flexible specification of the model allows us to use different distributions for each component. Here, the Poisson and negative binomial components were considered. Note that negative binomial components for certain multiplicities may tend to Poisson components when the dispersion parameter tends to ∞. In this case, to avoid numerical problems and for the sake of simplicity, Poisson components were used.

The 1-5-7-15-30-NB regression model was selected based on the BIC criterion. Its loglikelihood value was -95809.5, pointing to a marked improvement

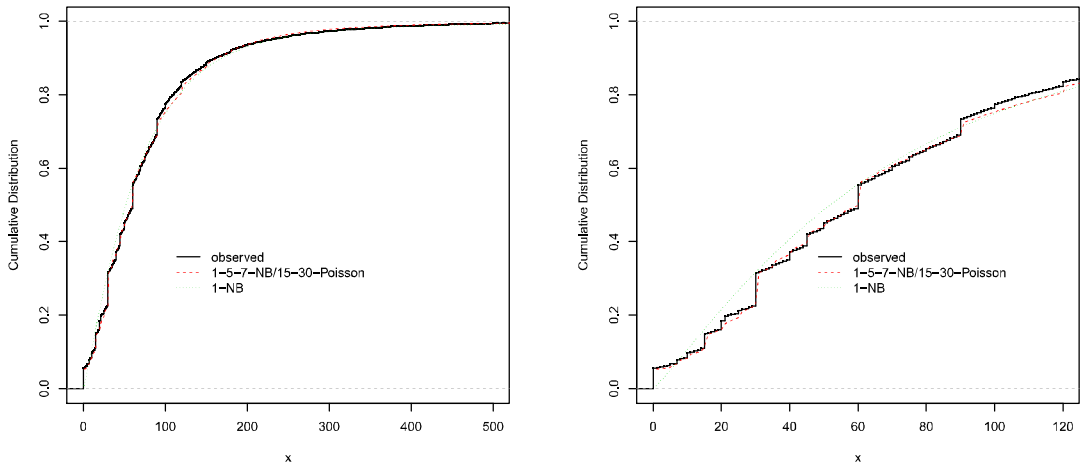

Figure 2: Empirical distribution function and the fitted cdfs of the different models

when using covariate information. Since the last two components presented a size parameter tending to ∞, indicative of Poisson rather than negative binomial components, the last two components with multiplicities of 15 and 30 days were fitted as Poisson (1-5-7-NB/15-30-Poisson regression model).

Figure 2 shows the empirical cumulative distribution function (cdf) of the data, together with the cdf for the fitted 1-5-7-NB/15-30-Poisson regression model and the simple 1-NB regression model. The right-hand plot is limited to 120 days for reasons of visibility. The spikes can be identified on the vertical jumps at certain points. The goodness of fit of the 1-5-7-NB/15-30-Poisson model is evident. The 1-5-7-NB/15-30-Poisson regression model captures the spikes in the data very accurately as is shown by the close resemblance between the estimated 1-5-7-NB/15-30-Poisson cdf and the empirical cdf. On the other hand, and as expected due to the special characteristics of the data, the 1-NB model performs poorly, failing to capture the spikes.

4.2. Parameter estimates

The results are shown in Table 4. Coefficient estimates and standard errors of the 1-5-7-NB/15-30-Poisson model are shown in the first six columns.

Standard errors were obtained on 1,000 non-parametric bootstrap replications. The last two columns show the p-value of the likelihood ratio test (LRT) statistic with the variable, and the marginal effect (ME) of the binary covariates.

	Coefficients (St.error) for each component						
	$m = 1$	$m = 5$	$m = 7$	$m = 15$	$m = 30$	LRT (Significance)	ME
Constant	3.739 (0.032)	1.181 (0.051)	−0.404 (0.64)	0.250 (0.156)	−0.119 (0.048)		
Gender	0.056 (0.015)	0.081 (0.037)	0.519 (0.341)	−0.079 (0.082)	0.080 (0.030)	< 0.001	8.59
Age	0.005 (0.005)	0.023 (0.009)	0.004 (0.084)	0.006 (0.021)	−0.030 (0.008)	< 0.001	
Driver	−0.058 (0.017)	0.05 (0.037)	0.057 (0.33)	−0.008 (0.079)	−0.041 (0.029)	< 0.001	8.41
Heavy vehicle	−0.032 (0.034)	0.001 (0.095)	0.016 (0.454)	−0.348 (0.268)	0.136 (0.063)	0.014	6.61
Injury score	0.578 (0.011)	0.82 (0.026)	1.232 (0.353)	0.545 (0.176)	0.679 (0.024)	< 0.001	
Mixing prop.	0.527 (0.005)	0.164 (0.004)	0.034 (0.003)	0.088 (0.005)	0.187 (0.005)		
NB size	2.296 (0.045)	4.063 (0.269)	0.774 (1188.14)				

Table 4: Results from fitting the 1-5-7-NB/15-30-Poisson regression model.

The LRT and ME columns in Table 4 help provide some insights into the relevance of the covariates for explaining the duration of sick leave of the motor accident victims. The LRT column captures the change in likelihood when the variable is included or excluded from the model. Note that the variable selection in the finite mixture regression setting is not simple, since each of the covariates appears in all the components. Coefficient estimates are probably correlated and, therefore, the interpretation of each coefficient estimate and its standard error taken individually need to be considered with some caution.

Analyzing the variables' aggregate explanatory capacity, by means of the LRT test, is more informative of the importance of each variable. The model was fitted by removing each covariate one at a time from all the components and the impact of this on the log-likelihood was analyzed. The results in Table 4 show that all the covariates were statistically significant. In terms of the variation level of the maximum likelihood value when the variable was

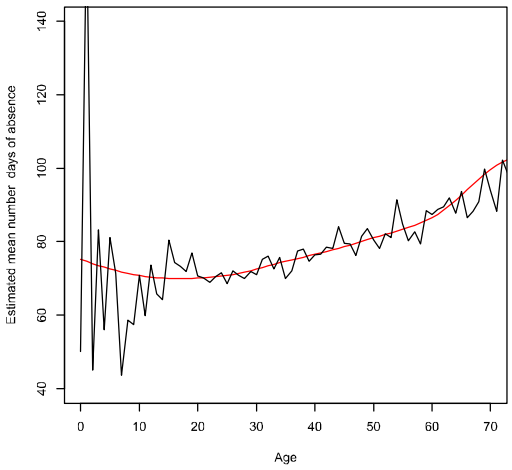

Figure 3: Smooth effect (red line) of age on the mean of the number of leave (black line) based on the fitted model.

removed, the variable capturing score of permanent injuries was, as expected, the most relevant, followed by the victim's age (values not shown).

The MEs of the binary covariates are also shown in Table 4. The ME of a dichotomous covariate was computed as the difference between the expected value of the period of sick leave when the covariate took the value 1 (presence) and when the covariate took the value 0 (absence). Sample mean values were considered for the rest of the covariates. Here, we can see that the impact on the expected duration of sick leave is similar for all binary regressors. When we observe any of these characteristics, the expected number of days of disability increases by less than ten in all cases.

Finally, the effect of continuous variables is analyzed. For the continuous covariates, the ME on the expected duration is illustrated in graph form. First, the effect of the victim's age on the expected duration of sick leave is displayed. Figure 3 shows the estimated duration of sick leave for different ages, smoothed with a LOESS smoother. The expected duration seems to increase exponentially with age, being relatively stable at around 70 days until the age 30 and, then, increasing quickly in line with age.

Now, the effect of the score for permanent injuries on the expected duration of sick leave is displayed in Figure 4. The expected duration clearly increases with permanent injury score.

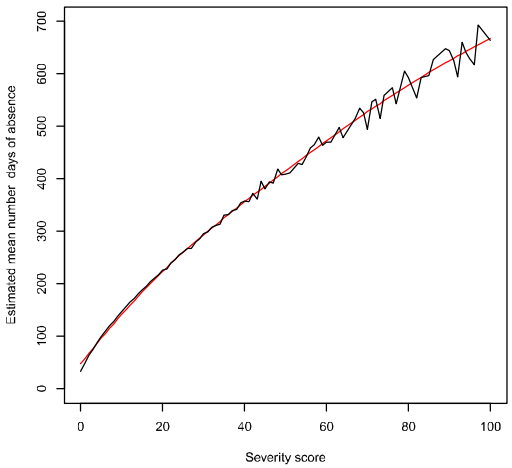

Figure 4: Smooth effect (red line) of injury score on the mean of the number of leave (black line) based on the fitted model.

We can observe that the score of permanent injuries has the greatest effect on the expected duration of sick leave. Now we split the data into victims with serious permanent injuries and victims without serious permanent injuries. A motor victim's permanent disability score over 10 according to the Spanish legislative scale is often considered by medical specialists to indicate that the motor victim has sustained serious permanent injuries. The serious injury indicator is constructed taking value one if the victim has a score for permanent injuries over 10 and 0 otherwise. Figure 5 shows the fitted model values for the sex, driver and heavy vehicle covariables, but here victims are grouped according to the serious injury indicator. Gray boxplots represent a serious injury indicator equal to 0; blue boxplots a serious injury indicator equal to 1. It is clear, as discussed above, that the role played by severity of permanent injuries is the most important. Likewise, we can see that the differences in the marginal effects conditional on the indicator of serious injuries are almost the same for the three variables. No interactions between the covariates are found.

In short, according to our results, males, drivers, crashes involving heavy vehicles and the elderly face longer periods of temporary sick leave to recover from their injuries after a motor accident. These results are in line with those reported in previous studies. Prior research suggests that male drivers are

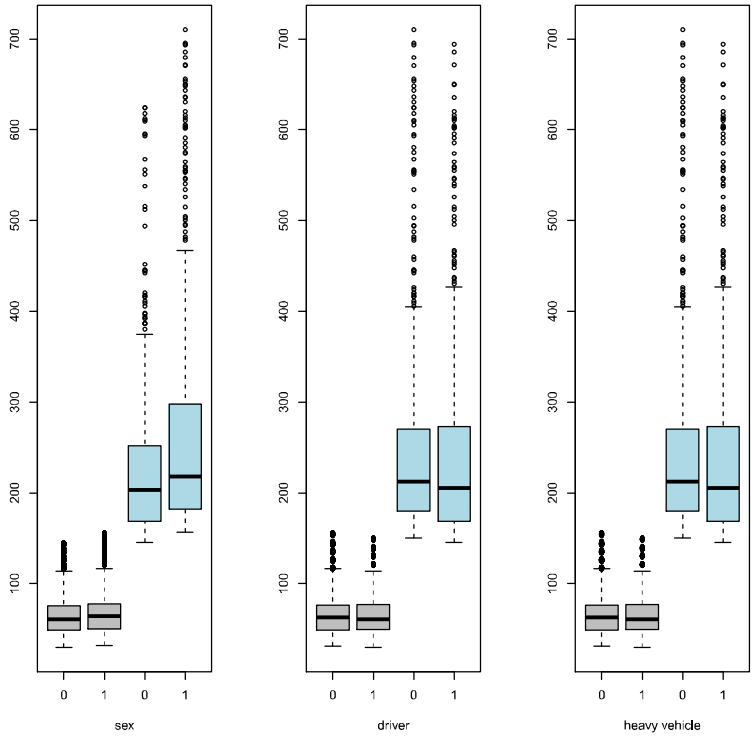

Figure 5: Duration of sick leave for individuals by gender, road user type and vehicle class differentiating by the serious injury covariate.

Note: Gray(blue) boxplots refer to victims without(with) serious permanent injuries.

involved in more serious crashes (Abdel-Aty, 2003; Tay and Rifaat, 2007), although other studies found that females are more likely to suffer severe injuries (Evans, 2001; Bedard et al., 2002). The identification of elderly victims as being a group at risk of sustaining serious injuries has been broadly recognized (Hanrahan et al., 2009; Delen et al., 2006; O´Donnell and Connor, 1996). The effect of the road user type on the severity of injury remains unclear. Vulnerable road users are generally associated with more serious injuries. Comparing between occupants of the vehicle, previous studies suggest that passengers are less likely to be seriously injury than drivers (Tay, 2016). It could be in part explained because those occupants occupying front seats are at a greater risk than those occupying rear seats (Smith and Cummings, 2004).

5. Concluding Remarks

Studies analyzing the repercussions of traffic accidents focus primarily on mortality and permanent injuries. Temporary injuries are rarely studied and, when they are, the focus is solely on the length of hospitalization, rather than on the total number of days the victim takes off sick as a result of the accident. This focus means that many of the consequences of road accidents are excluded from the analyses. In this article, we contribute to the study of temporary disabilities and the determination of their risk factors. More specifically, using a motor insurance dataset, the distribution of the number of days of sick leave taken by the victims of traffic accidents conditioned on a set of risk factors was analyzed.

The frequency distribution presented by the number of days of sick leave exhibited regular spikes at certain multiples, presumably reflecting the specific time scales used by doctors when determining the period of time they should allow to lapse before the next scheduled medical examination (in weeks, fortnights, months, etc.). This phenomenon is known in the literature as digit preference or the heaping of reported count data, i.e. rounding exact counts to even multiples of reported units. Studies dealing with digit preference/heaping assume that outcomes are indirect (or rounded) observations of a latent distribution and the principal idea underlying the approach is that of modeling the latent (unobserved) variable by smoothing techniques.

In the context dealt with here, we claim that valuable information would be lost with the application of smoothing techniques. This article has developed a suitable regression model for dealing with random variables that

present regular peaks. To account for digit preference/heaping, the regression model is based on finite mixtures of multiple discrete distributions. We found that the negative binomial-Poisson mixture regression model provided the best performance when working week, week, fortnight and month multiplicities were considered (1-5-7-NB/15-30-Poisson). This regression model specification captured the spikes in the data very accurately and the resemblance of its cumulative distribution function to the empirical distribution function was very similar to that obtained with standard alternatives, such as the more usual negative binomial regression model.

The analysis of the risk factors influencing the length of medical leave showed that characteristics such as gender, age, road user type, vehicle class and severity of the permanent injuries were statistically significant when explaining the expected duration of sick leave. The factor indicating serious permanent injuries was found to have the greatest impact on expected temporary disability. Gender, age, road user type and vehicle class had more moderate impacts in this regard. Males and drivers were associated with longer periods of temporary disabilities. The expected temporary disability of victims was also higher when heavy vehicles were involved. Finally, young victims were associated with shorter periods of temporary disabilities. The expected duration of temporary disability was relatively stable until victims entered their early thirties, but was increasing from this age on.

We argue that these results may be helpful in different areas. Decision makers may well be interested in modeling the data itself, i.e. the data with regular peaks, instead of smoothing the data to obtain the latent distribution, as digit preference or heaping literature assumes. Focusing on road traffic accidents, health planners, public managers and insurance providers may be interested in modeling the data with peaks when conducting a cost-benefit analysis. Different economic costs are derived from temporary disabilities of motor victims. Direct costs include motor insurance compensation payments, frequently related to a daily based compensation for the duration of the temporary disability, and payments for medical sick leave days, guaranteed by the Spanish labor legislation in case of victims are time off from work. These costs are highly related with the doctor's criteria to fix the medical leave time. For non-hospitalised motor victims, the evolution of their injuries can not be monitored by doctors on a real time basis. So, the fixation of a lapse of time between medical reexaminations is necessary.

Shortening the timing between medical examinations would enhance the accuracy of the estimation of the true victim's recovering time. However, a

more intensive use of medical services would be required and, therefore, more economic resources would be involved. In this cost-benefit analysis, public health planners and insurance managers may be interested in introducing an optimization criterion to fix the lapse of time between the medical examinations of traffic victims based on road accident information. The optimal lapse of time between reexaminations may be not the same for all motor victims and depend on the characteristics of the victim and the accident. In this framework the analysis of the impact of the risk factors on the shape of the distribution function of the length of medical leave would be required.

Acknowledgments. Research for this paper was initiated while the second author was visiting the Riskcenter Research Group at the University of Barcelona. The authors wish to acknowledge the support of the Spanish Ministry for grant ECO2015-66314-R and ECO2016-76203-C2-2-P.

References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34, 597–603.

Alemany, R., Ayuso, M., Guillén, M., 2013. Impact of road traffic injuries on disability rates and long-term care costs in spain. Accident Analysis & Prevention 60, 95–102.

Anastasopoulos, P. C., 2016. Random parameters multivariate tobit and zero-inflated count data models: Addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. Analytic Methods in Accident Research 11 (Supplement C), 17 – 32.

Ayuso, M., Bermúdez, L., Santolino, M., 2015. Modelización del tiempo de hospitalización en lesiones por tránsito. Salud Pública de México 57, 161 – 169.

Ayuso, M., Bermúdez, L., Santolino, M., 2016. Copula-based regression modeling of bivariate disability severity of temporary and permanent motor injuries. Accident Analysis & Prevention 89, 142–150.

Bedard, M., Guyatt, G., Stones, M., Hirdes, J., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis & Prevention 34, 717–727.

Berecki-Gisolf, J., Collie, A., McClure, R., 2013. Work disability after road traffic injury in a mixed population with and without hospitalization. Accident Analysis and Prevention 51, 129–134.

Bermúdez, L., Karlis, D., Santolino, M., 2017. A finite mixture of multiple discrete distributions for modelling heaped count data. Computational Statistics and Data Analysis 112 (Supplement C), 14 – 23.

Blincoe, L., Seay, A., Zaloshnja, A., Miller, T., Romano, E., Luchter, S., Spicer, R., 2002. The economic impact of motor vehicle crashes, 2000. Tech. rep., NHTSA Technical Report, DOT HS 809 446.

Boucher, J.-P., Santolino, M., 2010. Discrete distributions when modeling the disability severity score of motor victims. Accident Analysis & Prevention 42 6, 2041–2049.

Buitenhuis, J., de Jong, P., Jaspers, J., Groothoff, J., 2009. Work disability after whiplash: a prospective cohort study. Spine 34, 262–267.

Camarda, C. G., Eilers, P. H., Gampe, J., 2008. Modelling general patterns of digit preference. Statistical Modelling 8 (4), 385–401.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis & Prevention 38, 434–444.

Ebel, B., Mack, C., Diehr, P., Rivara, F., 2004. Lost working days, productivity, and restraint use among occupants of motor vehicles that crashed in the united states. Injury Prevention 10, 314–319.

Evans, L., 2001. Female compared with male fatality risk from similar physical impacts. Journal of Trauma 50, 281–288.

Gardner, R., Smith, G., Chany, A., Fernandez, S., McKenzie, L., 2007. Factors Associated With Hospital Length of Stay and Hospital Charges of Motor Vehicle Crash Related Hospitalizations Among Children in the United States. Archives of Pediatrics and Adolescent Medicine 161, 889–895.

Guria, J. C., 1990. Length of hospitalization. an indicator of social costs of disabilities from traffic injuries. Accident Analysis & Prevention 22 (4), 379 – 389.

Hanrahan, R., Layde, P., Zhu, S., Guse, C., Hargarten, S., 2009. The association of driver age with traffic injury severity in wisconsin. Traffic Injury Prevention 10, 361–367.

Lord, D., Washington, S. P., Ivan, J. N., 2005. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis & Prevention 37 (1), 35 – 46.

Mannering, F., Bhat, C., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research 1, 1–22.

Miller, R., Galbraith, M., 1995. Estimating the costs of occupational injury in the united states. Accident Analysis and Prevention 27 (6), 741–747.

O´Donnell, C., Connor, D., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accident Analysis and Prevention 28, 739–753.

Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis & Prevention 41 (4), 683 – 691.

Park, B.-J., Lord, D., Hart, J. D., 2010. Bias properties of bayesian statistics in finite mixture of negative binomial regression models in crash data analysis. Accident Analysis & Prevention 42 (2), 741 – 749.

Park, B.-J., Lord, D., Lee, C., 2014. Finite mixture modeling for vehicle crash data with application to hotspot identification. Accident Analysis & Prevention 71 (Supplement C), 319 – 326.

Peek-Asa, C., Yang, J., Ramirez, M., Hamann, C., Cheng, G., 2011. Factors affecting hospital charges and length of stay from teenage motor vehicle crash-related hospitalizations among united states teenagers, 2002–2007. Accident Analysis & Prevention 43 (3), 595 – 600.

Santolino, M., Bolancé, C., Alcañiz, M., 2012. Factors affecting hospital admission and recovery stay duration of in-patient motor victims in spain. Accident Analysis & Prevention 49, 512–519.

Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43, 1666–1676.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accident Analysis & Prevention 29 (6), 829 – 837.

Shibata, A., Fukuda, K., 1994. Risk factors of fatality in motor vehicle traffic accidents. Accident Analysis & Prevention 26 (3), 391 – 397.

Siegel, J. S., Swanson, D. A., 2004. The methods and materials of demography. Elsevier Academic Press, Amsterdam.

Smith, K., Cummings, P., 2004. Passenger seating position and the risk of passenger death or injury in traffic crashes. Accident Analysis & Prevention 36, 257–260.

Tay, R., 2016. Comparison of the binary logistic and skewed logistic (scobit) models of injury severity in motor vehicle collisions. Accident Analysis and Prevention 88, 52–55.

Tay, R., Rifaat, S., 2007. Factors contributing to the severity of intersection crashes. Journal of Advanced Transportation 41, 245–265.

Torelli, N., Trivellato, U., 1993. Modelling inaccuracies in job-search duration data. Journal of Econometrics 59 (1), 187–211.

Wang, B., Wertelecki, W., 2013. Density estimation for data with rounding errors. Computational Statistics & Data Analysis 65, 4–12.

Wang, H., Heitjan, D. F., 2008. Modeling heaping in self-reported cigarette counts. Statistics in medicine 27 (19), 3789–3804.

Wang, H., Shiffman, S., Griffith, S. D., Heitjan, D. F., 2012. Truth and memory: Linking instantaneous and retrospective self-reported cigarette consumption. The Annals of Applied Statistics 6 (4), 1689–1706.

Wolff, J., Augustin, T., 2003. Heaping and its consequences for duration analysis. Allgemeines Statistisches Archiv 87, 1–28.

Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accident Analysis and Prevention 59, 506–521.

Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Analytic Methods in Accident Research 1 (Supplement C), 39 – 52.