Discrete distributions when modelling the disability severity score in motor insurance claims

6th Conference in Actuarial Science and Finance

Miguel Santolino Jean-Phlipe Boucher

Department of Econometrics, RFA-IREA, University of Barcelona
msantolino@ub.edu
Département de Mathématiques, Université du Québec à Montréal
boucher.jean-philippe@uqam.ca

Samos, June 3-6, 2010

Contents

- Introduction
- Methodology
- Applications
- Conclusions

Motor Bodily injury (BI) claims: Spanish case

- BI claims have the largest impact on insurers' claims expenditure.
- A legislative compensation system, called *baremo*, is in force for claim settlements.
- The compensation stipulated comprises compensatory awards for non-pecuniary and pecuniary damages.
- The financial compensation is automatically fixed according to the severity of the injury, age and annual incomes.

How does the baremo work?

- Three concepts entitle to be compensated: death, temporary disability and permanent disability.
- A basic compensation amount is awarded for each one of these concepts. This compensation is stipulated to meet the non-pecuniary damages of the victim.
- Correction factors are applied to this basic compensation to compensate for pecuniary damages.

Claim settlement (I)

- Insurers usually seek to reach an amicable agreement with the plaintiff in order to settle BI claims as quickly as possible (i.e., interest rate payments, judicial expenses and so forth).
- The claim agreement is pursued when the victim is fully recovered. No discrepancies relating to the number of recovery days are expected.
- Neither disagreements as regards the annual income of the victim, age or year of settlement.
- Differences may appear in the evaluation of the permanent disability severity. Therefore, the most controversial issue in a compensation agreement is typically determining the severity score for permanent disability (the unique concept related to the future).

Claim settlement (II): Goals

- The underlying disability severity is modelled by means of a zero-altered regression models. This methodology provides probability estimates of severity scores for disability.

- The point estimate and the upper bound for the expected financial award for non-pecuniary damages that result from the disability are derived from the probability estimates of severity scores (practical applications: amount to offer in the negotiation, reserving purposes,).

Basic distributions

There are many discrete probability distributions that can be used to model the severity score of an injured motor victim. Let the response variable take the value y_i, which is the severity score for the permanent disability sustained by the $i-th$ motor victim resulting from a traffic accident.

Because the score is limited to values from 0 to 100, then the probability function of the severity score Z_i is $\Pr[Z_i = y_i] = \Pr[y_i] / \Pr[Y_i \leq 100]$, where y_i is a discrete variable.

Poisson distribution

The starting point for the modeling of a random variable is the Poisson distribution. The probability function of Y_i would be:

$$\Pr[Y_i = y_i] = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots, \tag{1}$$

where the $(k \times 1)$ vector of explanatory variables x is included in the model with the mean parameter $\lambda_i = \exp(x_i'\beta)$, being β the $(k \times 1)$ vector of coefficients.Here the Poisson distribution is extended by overdispersion, known as the generalized Poisson distribution $\Pr[Y_i = y_i] = \frac{\lambda_i(\lambda_i + (\varphi_i - 1)y_i)^{y_i - 1}}{y_i!} \varphi_i^{-y_i} \exp(-\frac{\lambda_i + (\varphi_i - 1)y_i}{\varphi_i})$. where the dispersion function φ_i is modelled as $\varphi_i = 1 + \exp(z_i'\gamma)$.

Geometric distribution

A more intuitive distribution is the Geometric distribution. This distribution is known to model the *number of successes before a single failure*.

Applied to the severity score, the justification of the Geometric distribution is as follows. To obtain a specific severity score y_i, the i-*th* victim must have all the symptoms associated with values $1, 2, \ldots, y_i$, which we call *successes*, without having all the symptoms of score $y_i + 1$, which represents a *failure*. Using this interpretation, the score probability can be expressed as:

$$\Pr[Y_i = y_i] = p_i^{y_i}(1 - p_i), \tag{2}$$

where covariates can be included in the model using a logit function, so $p_i = \exp(x_i'\beta)/(\exp(x_i'\beta) + 1)$.

Negative Binomial distribution

The Negative Binomial (NB) distribution may be constructed as a generalization of the Geometric distribution, where, instead of modeling the *number of successes before a single failure*, it models the *number of successes before a specific quantity of failures*. The probability distribution is:

$$\begin{aligned}\Pr[Y_i = y_i] &= \tfrac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\tfrac{\lambda_i}{\alpha^{-1}+\lambda_i}\right)^{y_i}\left(\tfrac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)^{\alpha^{-1}} \\ &= \tfrac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}p_i^{y_i}(1-p_i)^{\alpha^{-1}}\end{aligned} \tag{3}$$

where the $\lambda_i = \exp(x_i'\beta)$ and $\Gamma(\cdot)$ is the gamma function. The second equality highlights the fact that the logit transformation of the regressors of the Geometric distribution is now generalized by $p_i = \left(\frac{\lambda_i}{\alpha^{-1}+\lambda_i}\right)$.

Zero-altered models

The score reflects the degree of severity of the permanent disability, thus motor victims with only temporary disabilities resulting from the accident have a score equal to zero. Based on our intuition that accidents involving permanent disabilities may not have the same characteristics as those involving only temporary disabilities, zero-altered distributions are also considered. In this section, we analyze zero-inflated and hurdle models based on the three previous basic distributions.

Zero-inflated models

The idea of zero-inflated (ZI) models is to use a finite mixture model combining an indicator distribution for the zero case and a basic discrete distribution. Consequently, this distribution will account for the excess of zeros. The density of this kind of model, with $0 < \phi_i < 1$, can be expressed as:

$$P(Y_i = y_i) = \begin{cases} \phi_i + (1-\phi_i)Pr(K_i = 0) & \text{for } y_i = 0 \\ (1-\phi_i)Pr(K_i = y_i) & \text{for } y_i = 1, 2, \ldots \end{cases}$$

where the random variable K follows a basic distribution. The $(p \times 1)$ vector of regressors w is included such as $\phi_i = \exp(w_i'\gamma)/(\exp(w_i'\gamma)+1)$, being γ the $(p \times 1)$ vector of coefficients.

Hurdle models

A different approach to modify a basic discrete distribution is to use it as a part of a two-process distribution.The first part of the model is a binary outcome model and the second part is a discrete distribution that takes the values $\{1, 2, 3, \ldots\}$. Consequently, in the modeling of the second part, a choice between a basic discrete distribution (truncated or shifted) and a discrete distribution with support domain $\{1, 2, 3, \ldots\}$ must be made. Let $f_{i,1}(\cdot)$ and $f_{i,2}(\cdot)$ be two probability mass functions with respective support $\{0, 1\}$ and $\{0, 1, \ldots\}$ depending on parameter vectors θ_1 and θ_2. The random variable Y_i obeys the hurdle distribution if:

$$P(Y_i = y) = \begin{cases} f_{i,1}(0) & \text{for } y = 0 \\ \frac{1-f_{i,1}(0)}{1-f_{i,2}(0)} f_{i,2}(y) = \Psi_i f_{i,2}(y) & \text{for } y = 1, 2, \ldots \end{cases},$$

where $\Psi_i = \frac{1-f_{i,1}(0)}{1-f_{i,2}(0)}$.

Zero-inflated and hurdle models models can be expressed as a compound sum of two random variables:

$$Y = \sum_{i=1}^{M} X_i.$$

where X_i are i.i.d., independent from M, and $Y = 0$ if $M = 0$. Under this construction, there are two possibilities:

- For the ZI distribution: $M \sim Bernoulli(\phi_i)$ with X_i taking values $0, 1, 2, 3, ...$
- For the hurdle distribution: $M \sim Bernoulli(\delta_i)$ with X_i taking only positive values $1, 2, 3, ...$

A main difference between the two distributions is thus in the way a zero score is obtained. For the hurdle distribution, it happens only if $M = 0$, while for the ZI model it happens if $M = 0$ or if $M = 1$ and $X_1 = 0$.

The conditional moment (CM) test proposed by Santos-Silva and Windmeijer (2001) can be adapted to check if the hurdle specification is valid. When the *separation hypothesis* assumption of the hurdle model is fulfilled, the equality $E\big[Y - E[M]E[X]\big] = 0$ is satisfied. This equality may be tested using the CM test described by Newey (1985) and Tauchen (1985), and explained in Cameron and Trivedi (1998).

Database: big sample

The database consists of a random sample of $18,363$ non-fatal victims. All of the victims needed at least one day to recover from the injuries caused by the accident. The sample covers all the provinces of Spain. All victims were compensated for their injuries in 2007. The at fault system is in place in Spain.

The dependent variable to model is the variable *score*. The final score is stated by judicial decision, or agreed upon between parties.

Figure 1. Histogram for the disability severity score

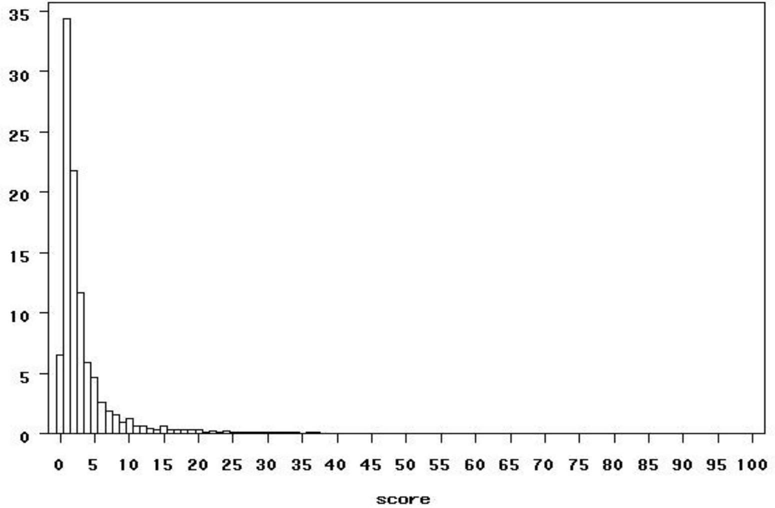


Table 1. Variable description and some statistics

Variable	Description	Mean	Std.Dev.
score	Severity score of permanent disability	3.930	6.510
gend	1 if victim is male.	0.451	0.497
age	Age of victim.	38.251	17.029
driv	1 if victim was the driver	0.482	0.500
pas	1 if victim was the passenger	0.383	0.486
pedcy	1 if victim was either a pedestrian or a cyclist	0.134	0.341
hrd	Number of recovering days in hospital (in log.).	0.341	0.891
ird	Number of recovering days with inability (in log.).	3.900	1.254
nird	Number of recovering days without inability (in log.).	2.026	1.972

Model comparison

Classical hypothesis tests can be perfomed to accept or reject nested models. Application of these tests on score data for the α parameter of the NB distribution leads to the rejection of Poisson and Geometric distributions in favor of the Negative Binomial for basic (both *p-values* less than 0.001), zero-inflated (both *p-values* less than 0.001), and hurdle (both *p-values* less than 0.001) constructions.

We are still undecided about how a severity score of zero is generated. Three candidates remain in the modeling of the severity score: the basic, the zero-inflated and the hurdle Negative Binomial regressions.

Table 2 The NB, the ZI-NB and Hurdle-NB models

	NB	ZI-NB	Hurdle-NB
Loglikelihood	39,233.84	-39,166.13	-36,770.41
AIC	78,484	78,357	73,569
BIC	78,546	78,451	73,678

N=18,363

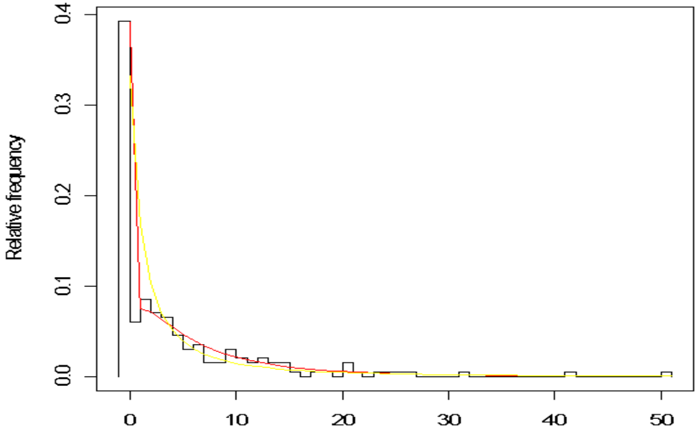
Akaike's and Bayesian Information Criterion (AIC and BIC) clearly gives an advantage to the hurdle NB model. To analyze if the observed differences in the log-likelihood and the information criterion are statistically significant, a test based on the difference in the log-likelihoods can be performed. Indeed, for independent observations, a log-likelihood ratio test for non-nested models, developed by Vuong (1989), can be used to see whether the hurdle NB model is statistically better than the zero-inflated NB and the NB model. Applied to the data, the Vuong test shows that the information criterion of the hurdle NB model is statistically different from the other models, with *p-values* of less than 0.001 for each test. Therefore, from a statistical viewpoint, the hurdle NB model offers the highest fit.

The CM test rejected the null hypothesis (*p-value* of less than 0.001). it is well-known that CM tests are very powerful when used with many observations (here, $n = 18,363$), meaning that the null hypothesis is usually rejected.

Another explanation lies in the fact that it is possible that common unidentified individual characteristics, i.e. heterogeneity, affect both processes of the hurdle model.

Database II: small sample

Sample of 180 claims settled by Court decision between 2001 and 2003. It covers mainly Catalonia-Aragón.



The ZIGP distribution was the preferred method. The ZIGP distribution is a mixture of a Bernoulli distribution and a generalized Poisson distribution.Here the Poisson distribution is extended by overdispersion and zero-inflation parameters, known as the ZIGP distribution. Count data with a large zero fraction and a heavy tail are common in a number of applications.

A) Claim negotiation

The point estimate of the expected compensation awarded by the Courts and the upper bound can be derived from,

$$E[y_i * Cpp_{ij}] = \sum_{h=1}^{100} \Pr[y_i = h] * h * Cpp_{ij|Y_i=h}$$
$$Var(y_i * Cpp_{ij}) = E[(y_i * Cpp_{ij})^2] - (E[y_i * Cpp_{ij}])^2$$

where Cpp_{ij} is the financial compensation per point (depends on the settlement year j, the victim's age and the total score y_i).This estimate would be the amount to offer in the negotiation process and the upper-bound estimate the maximum amount can be accepted by the insurer in order to avoid legal proceedings.

B) Claim reserving

Statistical methods based on individual claim information have grown in importance in recent years. In the table, the total compensation awarded by the insurer is compared with the sum of financial compensations estimate.

	Total amount (in euros)	$\frac{\text{Estimated provision}}{\text{Empirical compensation}}$
Empirical compensations	631294.10	-
A) Estimated claims provision	691922.52	109.60%

Conclusions

- The expected size of BI claim costs will have implications for compensation negotiation and individual reserves.
- In the Spain it can be reduced to an estimate of the victim disability severity, since remaining factors are known at the time of settlement.
- The main advantage of modelling injury severity rather than directly modelling the financial compensation is that financial effects are withdrawn from the motor accident BI claim evaluation. Injury severity does not depend on economic factors such as the settlement year, the inflation rate or the cost of medical services, among others.It allows insurance companies to monitor the real severity level underlying the claim.
- This methodology may be accomodated to other European States that also apply disability scales (France, Italy, Portugal, Belgium, etc.)

References

- Ayuso, M., Santolino, M. 2007. Predicting automobile claims bodily injury severity with sequential ordered logit models. IME, 4, 71-83.
- Baughman, A.L. 2007. Mixture model framework facilities understanding of zero-inflated and hurdle models for zount data. Journal of Biopharmaceutical Statistics, 17, 943-946.
- Boucher, J.-P., Denuit, M., Guillén, M. 2007. Risk classification for claim counts: A comparative analysis of various zero-inflated mixed Poisson and hurdle models. NAAJ,11, 110–131.
- Boucher, J.-P., Denuit, M., Guillén, M. 2009. Number of accidents or number of claims? an approach with zero-inflated Poisson models for panel data. JRI, 76, 821-846.
- Czado et al. 2007. Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. Statistical Modelling, 7, 125-153.

- Lambert, D. 1992. Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics, 34, 1–14.
- Lord, D., Washington, S., Ivan, J. 2005. Poisson, Poisson-Gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. AAP, 37, 35-46.
- Mullahy, J. 1986. Specification and testing in some modified count data models. Journal of Econometrics, 33, 341–365.
- Newey, W. 1985. Generalized method of moment specification testing. Journal of Econometrics, 29, 229–256..
- Santos-Silva, J., Windmeijer, F. 2001. Two-part multiple spell models for health care demand. Journal of Econometrics, 104, 67–89.
- Santolino, M., Boucher, J.-P. 2010. Modelling the Disability Severity Score in Motor Insurance Claims: An Application to the Spanish case. JFDM, 6,2.
- Vuong, Q. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57, 307–333.