Computational Statistics and Data Analysis 56 (2012) 3988–3999

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking

Lluís Bermúdez [a,∗], Dimitris Karlis [b]

[a] Risc en Finances i Assegurances-IREA, Universitat de Barcelona, Spain

[b] Athens University of Economics and Business, Greece

ARTICLE INFO

Article history:
Received 13 July 2011
Received in revised form 15 May 2012
Accepted 15 May 2012
Available online 2 June 2012

Keywords:
Zero-inflation
Overdispersion
EM algorithm
Automobile insurance
A priori ratemaking

ABSTRACT

Bivariate Poisson regression models for ratemaking in car insurance have been previously used. They included zero-inflated models to account for the excess of zeros and the overdispersion in the data set. These models are now revisited in order to consider alternatives. A 2-finite mixture of bivariate Poisson regression models is used to demonstrate that the overdispersion in the data requires more structure if it is to be taken into account, and that a simple zero-inflated bivariate Poisson model does not suffice. At the same time, it is shown that a finite mixture of bivariate Poisson regression models embraces zero-inflated bivariate Poisson regression models as a special case. Finally, an EM algorithm is provided in order to ensure the models' ease-of-fit. These models are applied to an automobile insurance claims data set and it is shown that the modeling of the data set can be improved considerably.

1. Introduction

In a recent paper Bermúdez (2009) describes bivariate Poisson (BP) regression models for ratemaking in car insurance. The central idea is that the dependence between two different types of claim must be taken into account to achieve better ratemaking. BP regression models are presented, therefore, as an instrument that can account for the underlying correlation between two types of claim arising from the same policy (i.e. third-party liability claims and all other automobile insurance claims). The paper concludes that even when there are small correlations between the claims, major differences in ratemaking can nevertheless appear. Thus, using a BP model results in ratemaking that has larger variances and, hence, larger loadings in premiums than those obtained under the independence assumption.

The paper also includes zero-inflated bivariate Poisson (ZIBP) models so as to inflate the $(0, 0)$ cell and to account for the excess of zeros and overdispersion typically observed in this type of data set. This produces the best goodness of fit among the bivariate Poisson models considered. In conclusion, the independence assumption should be rejected when using either BP or ZIBP regression models, but one question still remains unresolved: do ZIBP models constitute the best option for dealing with the unobserved heterogeneity usually observed in such a data? The aim of the present paper is to examine this question further by considering alternative bivariate models that might account for both these features of the data, i.e. the excess of zeros and overdispersion.

In the context of automobile insurance, the problem of unobserved heterogeneity is caused by the differences in driving behavior among policyholders that cannot be observed or measured by the actuary, such as a driver's reflexes, his or her aggressiveness, or knowledge of the Highway Code, among others. The main consequence of unobserved heterogeneity is

∗ Correspondence to: Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Diagonal 690, 08034-Barcelona, Spain. Tel.: +34 93 4034853; fax: +34 93 4034892.

E-mail address: lbermudez@ub.edu (L. Bermúdez).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.05.016

overdispersion, i.e. data exhibiting variance larger than mean, which it invalidates the use of a simple Poisson distribution. The presence of excess of zeros in most insurance data sets can be also seen as a consequence of unobserved heterogeneity.

In the univariate case, Lambert (1992) introduced the zero-inflated Poisson regression model. Since then, there has been a considerable increase in the number of applications of zero-inflated regression models based on several different distributions. A comprehensive discussion of these applications can be found in Winkelmann (2008) and a specific application to insurance ratemaking is addressed in Boucher et al. (2007). Zero inflated negative binomial regression models have been also described as for example in Wang (2003) and Garay et al. (2011). See again Winkelmann (2008) for a description of a variety of such models and Denuit et al. (2007) for an exhaustive review of the models used in ratemaking systems for automobile insurance.

In the bivariate (or multivariate) case, the literature analyzing the excess of zeros and overdispersion is less developed. For example, zero-inflation in the bivariate case is examined in Gurmu and Elder (2008) and Karlis and Ntzoufras (2003) and the references therein, while in the multivariate case it is analyzed in Li et al. (1999). Recently, in the actuarial literature and for ratemaking purposes, Bermúdez (2009) and Bermúdez and Karlis (2011) deal with the bivariate and multivariate versions of the zero-inflated Poisson regression models, respectively. They tackle overdispersion via the excess of zeros, i.e. zero-inflated models.

A natural approach for accounting for overdispersion is to consider models with some overdispersed marginal distribution, as opposed to bivariate Poisson models. In this paper we consider an m-finite mixture of bivariate Poisson regressions (m-FMBP) extending the no-covariate cases presented in Karlis and Meligkotsidou (2007). This model has a number of interesting features: first, the zero-inflated model represents a special case; second, it allows for overdispersion; and, third, it allows for an elegant interpretation based on the typical clustering application of finite mixture models. To the best of our knowledge, this model is new to the literature, so in what follows we seek to explain its properties as well as to discuss appropriate estimation approaches.

The rest of of the paper proceeds as follows. The new model is described in the next section, followed by the development of an EM algorithm for parameter estimation. The model is then applied to the same data set as in Bermúdez (2009). Finally, we conclude with some remarks.

2. The proposed model

2.1. A bivariate Poisson distribution

Consider random variables X_k, $k = 1, 2, 3$ which follow independent Poisson distributions with parameters $\lambda_k \geq 0$, respectively. Then the random variables $Y_1 = X_1 + X_3$ and $Y_2 = X_2 + X_3$ jointly follow a bivariate Poisson distribution, denoted as $BP(\lambda_1, \lambda_2, \lambda_3)$, with joint probability function given by

$$\begin{aligned} P_{Y_1,Y_2}(y_1, y_2) &= P(Y_1 = y_1, Y_2 = y_2) \\ &= BP(y_1, y_2; \lambda_1, \lambda_2, \lambda_3) \\ &= e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{s=0}^{\min(y_1,y_2)} \binom{y_1}{s} \binom{y_2}{s} s! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^s . \end{aligned}$$

The above bivariate distribution allows for dependence between the two random variables. It is also related to a common shock model. Marginally each random variable follows a Poisson distribution with $E(Y_1) = \lambda_1 + \lambda_3$ and $E(Y_2) = \lambda_2 + \lambda_3$. Moreover, $\text{Cov}(Y_1, Y_2) = \lambda_3$, and hence λ_3, is a measure of dependence between the two random variables. If $\lambda_3 = 0$ then the two variables are independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions (also known as a double Poisson distribution). For a comprehensive treatment of the bivariate Poisson distribution and its multivariate extensions the reader is referred to Kocherlakota and Kocherlakota (1992) and Johnson et al. (1997).

For greater flexibility, we can assume a bivariate Poisson regression model where each of the parameters of the BP is related to some covariates through a log link function, i.e. by assuming

$$\log \lambda_{ki} = \boldsymbol{\beta}_k^T \boldsymbol{x}_{ki}, \quad k = 1, 2, 3,\ i = 1, \dots, n,$$

where $\boldsymbol{x}_{ki}$ is a vector of covariates for the i-th observation related to the k-th parameter and $\boldsymbol{\beta}_k$ is the associated vector of regression coefficients. Note that $\boldsymbol{x}$ does not need to be the same for all the parameters. Likewise note that according to Karlis and Ntzoufras (2003), it is perhaps a good idea not to use the same covariates in all the parameters since this may lead to problems in their interpretation. For example, since the marginal mean for Y_1 is $\lambda_1 + \lambda_3$ using the same covariates in both may create problems of interpretation especially if the signs of the regression coefficients differ. R package `bivpois` can be used to fit this model based on an EM algorithm.

In this model, and since the marginal distributions are Poisson, the marginal means and variances are equal, moreover, the correlation is positive. Therefore, there we need to consider extensions to allow for overdispersion (variance greater than the mean) and a possible negative correlation.