Insurance: Mathematics and Economics 55 (2014) 225–249

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

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

Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape

Nadja Klein [a], Michel Denuit [b,*], Stefan Lang [c], Thomas Kneib [a]

[a] University of Göttingen, Germany
[b] Université Catholique de Louvain, Belgium
[c] University of Innsbruck, Austria

ARTICLE INFO

Article history:
Received October 2013
Received in revised form
December 2013
Accepted 2 February 2014

Keywords:
Overdispersed count data
Mixed Poisson regression
Zero-inflated Poisson
Negative binomial
Zero-adjusted models
MCMC
Probabilistic forecasts

ABSTRACT

Generalized additive models for location, scale and, shape define a flexible, semi-parametric class of regression models for analyzing insurance data in which the exponential family assumption for the response is relaxed. This approach allows the actuary to include risk factors not only in the mean but also in other key parameters governing the claiming behavior, like the degree of residual heterogeneity or the no-claim probability. In this broader setting, the Negative Binomial regression with cell-specific heterogeneity and the zero-inflated Poisson regression with cell-specific additional probability mass at zero are applied to model claim frequencies. New models for claim severities that can be applied either per claim or aggregated per year are also presented. Bayesian inference is based on efficient Markov chain Monte Carlo simulation techniques and allows for the simultaneous estimation of linear effects as well as of possible nonlinear effects, spatial variations and interactions between risk factors within the data set. To illustrate the relevance of this approach, a detailed case study is proposed based on the Belgian motor insurance portfolio studied in Denuit and Lang (2004).

1. Introduction

Calculations of motor insurance premiums are based on detailed statistical analyses of large data bases maintained by insurance companies, recording individual claim experience. The actuarial evaluation relies on a statistical model incorporating all the available information about the risk. Premiums then often vary by the territory in which the vehicle is garaged, the use of the vehicle (driving to and from work or business use, for example) and individual characteristics (such as age, gender, occupation and marital status of the main driver of the vehicle, for instance). If the policyholders misrepresent any of these classification variables in their declaration, they are subject to loss of coverage when they are involved in a claim. There is thus a strong incentive for accurate reporting of risk characteristics making insurance data reliable.

It is now common practice to achieve a priori risk classification with the help of Generalized Linear Models (GLMs). See, e.g., Denuit et al. (2007) for an introduction in relation with motor insurance. GLMs are so called because they generalize the classical linear model based on the Normal distribution to similar regression models for Poisson, Binomial, Gamma or Inverse-Gaussian responses, for instance. The main drawback of GLMs is that covariate effects are modeled in the form of a linear predictor. This is not a problem for categorical explanatory variables coded by means of binary variables, but a strong restriction for continuous explanatory variables which may have a nonlinear effect on the score. It has been common practice in insurance companies to model possibly nonlinear effects by means of polynomials. However, it is now well documented that low-degree polynomials are often not flexible enough to capture the variability in the data and that increasing their degree produces unstable estimates, especially for extreme values of the covariates. Although banding results in a loss of information, a model employing a banded version of a continuous covariate is sometimes considered more practical than one which employs the (untransformed) continuous variable. However, there is no general rule to determine the optimal choice of cut-offs so that banding may bias risk evaluation.

Among continuous covariates, geographic area plays a particular role. It can either be seen as a function of two coordinates if exact locations are available or a function of an administrative areal variable if spatial information is aggregated. In any case, actuaries wish to estimate the spatial variation in risk premium and to

* Corresponding author. Tel.: +32 10472835.
E-mail address: michel.denuit@uclouvain.be (M. Denuit).

http://dx.doi.org/10.1016/j.insmatheco.2014.02.001
0167-6687/

price accordingly. Spatial postcode methods for insurance rating attempt to extract information which is in addition to that contained in standard factors (like age or gender, for instance). With the regression models discussed in the present paper, the effect of continuous and spatial covariates is modeled on the score scale by means of smooth, unspecified functions estimated from the data.

Generalized additive models (GAMs) as developed in Hastie and Tibshirani (1990) and popularized by Wood (2006) provide a convenient framework to overcome the linearity assumptions inherent to GLMs when smooth effects of continuous covariates need to be included in an additive predictor. Inference can be realized by cross validation as in Wood (2004), by mixed model representations as in Ruppert et al. (2003), Fahrmeir et al. (2004) and Wood (2008) or by Markov chain Monte Carlo (MCMC) simulations as in Brezger and Lang (2006), Jullion and Lambert (2007) and Lang et al. (2013).

The framework of generalized additive models for location, scale and shape (GAMLSS) introduced by Rigby and Stasinopoulos (2005) allows to extend GAMs to more complex response distributions where not only the expectation but multiple parameters are related to structured additive predictors with the help of suitable link functions. Structured additive regression relies on a unified representation of different model terms like parametric linear effects, smooth nonlinear effects of continuous covariates, interaction terms based on varying coefficients and spatial effects (Fahrmeir et al., 2013; Brezger and Lang, 2006). In particular, zero-inflated, skewed and zero-adjusted distributions can be embedded in this framework as special cases where all occurring parameters are related to regression predictors and may depend on a complex covariate structure. All these model terms rely on a unifying representation based on non-standard basis function specifications in combination with quadratic penalties in a frequentist formulation or Gaussian priors in a Bayesian approach.

In this broader setting, the Poisson assumption for claim frequencies made in Denuit and Lang (2004) is replaced with a mixed Poisson one, with Gamma distributed random effect (yielding the Negative Binomial distribution with cell-specific heterogeneity) or Bernoulli distributed random effect (yielding the zero-inflated Poisson distribution with cell-specific additional probability mass at zero).

In addition to claim frequencies, we also consider regression models for claim severities. Much attention has been paid in the actuarial literature to find suitable distributions to model claim sizes; see for example Klugman et al. (2004). Whereas Denuit and Lang (2004) studied claim frequencies and claim severities separately, we consider in this paper the so-called zero-adjusted models that allow to account for zeros in the analysis of the amount of loss directly without resorting to models for claim frequencies. Zero-adjusted distributions combine a continuous distribution on the positive real line and a point mass at zero, such that the probabilities for a claim and quantiles of the claim size distribution can be estimated in one model. Zero-adjusted models are in the line of Jørgensen and Paes de Souza (1994) and Smyth and Jørgensen (2002) where the zero claims are included using the Tweedie distribution. However, this model has the disadvantage that the probability at zero cannot depend on covariates whereas here, this key actuarial indicator is allowed to vary according to risk characteristics.

To select an appropriate response distribution and to specify several predictors that correspond for instance to variance, skewness or overdispersion of the distribution, we rely mainly on the deviance information criterion (DIC) of Spiegelhalter et al. (2002) whose performance in Bayesian count data regression within the framework of GAMLSS has been tested in Klein et al. (2013a). The choice of the distribution will be supported by normalized quantile residuals (Dunn and Smyth, 1996) and proper scoring rules (Gneiting and Raftery, 2007).

We highlight the advantages of complex Bayesian count data, skewed and zero-adjusted regression models for insurance claims data with a detailed analysis of a Belgian data set with more than 160,000 policies. Specifically,

- we consider the Poisson, zero-inflated Poisson and Negative Binomial regression models for claim frequencies, where suitable predictors are specified for the expected number of claims as well as for the probability of the structural zeros in the zero-inflated Poisson distribution and for the scale parameter of the Negative Binomial distribution.
- for claim severities, we extend the continuous models to zero-adjusted versions of the Gamma, Inverse-Gaussian and LogNormal distributions, we estimate the corresponding location and scale or shape parameters as well as the probability of a claim in terms of relevant covariates in an additive fashion.
- inference in all model formulations is based on iteratively weighted least squares approximations to the full conditionals in MCMC simulation techniques as suggested in Gamerman (1997) or Brezger and Lang (2006) and extended to the general framework of GAMLSS by Klein et al. (2013b).
- we benefit from a numerically efficient implementation in the free open software BayesX.
- compared to frequentist GAMLSS, the approach adopted here directly includes the choice of smoothing parameters in the estimation run and provides valid confidence intervals which are difficult to obtain from asymptotic maximum likelihood theory.

Our approach to zero-inflated, skewed and zero-adjusted models has therefore the full flexibility in the parametric distribution assumption. The structured additive modeling of all parameters allows to focus on specific aspects of the data that go beyond the mean. In particular,

- the separate modeling of the probability mass at zero as a function of the observable characteristics allows for an accurate analysis of this key actuarial indicator.
- cell-specific residual heterogeneity in the Negative Binomial model allows for more accurate risk predictions when deriving the predictive distributions of future claims.
- zero-adjusted models for the annual claim amounts are in accordance with the individual model of risk theory so that the actuarial analysis benefits from the numerous tools developed in that setting.

The claim frequencies models considered in the present paper have already been applied to insurance data. See, e.g., Yip and Yau (2005) or Boucher et al. (2007). However, previous applications of Negative Binomial or zero-inflated Poisson regression models to insurance data only allowed for linear effects of the covariates or applied preliminary banding techniques to transform continuous covariates into categorical ones. Zero-adjusted Gamma and Inverse-Gaussian models have been proposed by Heller et al. (2006), Bortoluzzo et al. (2011) and Resti et al. (2013) but their analysis only allowed for linear effects of the covariates, too. See also Heller et al. (2007) for a related model extending the Tweedie construction beyond the Poisson–Gamma setup. The present paper innovates in that nonlinear effects are allowed using the efficient inference techniques developed by Klein et al. (2013a). The effect of continuous covariates on the score are quantified by means of unknown smooth functions that do not need to be specified a priori under parametric form but are estimated directly from the data.

The remainder of the paper is organized as follows. Section 2 is devoted to the presentation of the data set used in our study. In Section 3 we introduce the specification of Bayesian models for analyses of claim frequencies as well as severities. We describe the underlying inference and we give guidelines for model choice.