Available online at www.sciencedirect.com
ScienceDirect
Insurance: Mathematics and Economics 40 (2007) 58–76
INSURANCE MATHEMATICS & ECONOMICS
www.elsevier.com/locate/ime

Actuarial statistics with generalized linear mixed models

Katrien Antonio*, Jan Beirlant

University Center for Statistics, KU Leuven, W. de Croylaan 54, 3001 Heverlee, Belgium

Received March 2005; received in revised form February 2006; accepted 23 February 2006

Abstract

Over the last decade the use of generalized linear models (GLMs) in actuarial statistics has received a lot of attention, starting from the actuarial illustrations in the standard text by McCullagh and Nelder [McCullagh, P., Nelder, J.A., 1989. Generalized linear models. In: Monographs on Statistics and Applied Probability. Chapman and Hall, New York]. Traditional GLMs however model a sample of independent random variables. Since actuaries very often have repeated measurements or longitudinal data (i.e. repeated measurements over time) at their disposal, this article considers statistical techniques for modelling such data within the framework of GLMs. Use is made of generalized linear mixed models (GLMMs) which model a transformation of the mean as a linear function of both fixed and random effects. The likelihood and Bayesian approaches to GLMMs are explained. The models are illustrated by considering classical credibility models and more general regression models for non-life ratemaking in the context of GLMMs. Details on computation and implementation (in SAS and WinBugs) are provided.

Keywords: Non-life ratemaking; Credibility; Bayesian statistics; Longitudinal data; Generalized linear mixed models

1. Introduction

Over the last decade generalized linear models (GLMs) have become a common statistical tool for modelling actuarial data. Starting from the actuarial illustrations in the standard text by McCullagh and Nelder (1989), over applications of GLMs in loss reserving, credibility and mortality forecasting, a whole scala of actuarial problems can be enumerated where these models are useful (see Haberman and Renshaw (1996), for an overview). The main merits of GLMs are twofold. Firstly, regression is no longer restricted to normal data, but extended to distributions from the exponential family. This enables appropriate modelling of, for instance, frequency counts, skewed or binary data. Secondly, a GLM models the additive effect of explanatory variables on a transformation of the mean, instead of the mean itself.

Standard GLMs require a sample of independent random variables. In many actuarial and general statistical problems however the assumption of independence is not fulfilled. Longitudinal, spatial or (more general) clustered data are examples of data structures where this assumption is doubtful. This paper puts focus on repeated measurements and, more specific, longitudinal data, which are repeated measurements on a group of 'subjects' over time. The interpretation of 'subject' depends on the context; in our illustrations policyholders and groups of

* Corresponding author. Tel.: +32 (0) 16 32 67 69.
E-mail address: Katrien.Antonio@econ.kuleuven.be (K. Antonio).

0167-6687/$ - see front matter
doi:10.1016/j.insmatheco.2006.02.013

policyholders (risk classes) are considered. Since they share subject-specific characteristics, observations on the same subject over time often are substantively correlated and require an appropriate toolbox for statistical modelling.

Two popular extensions of GLMs for correlated data are the so-called marginal models based on generalized estimating equations (GEEs) on the one hand and the generalized linear mixed models (GLMMs) on the other hand. Marginal models are only mentioned indirectly and do not constitute the main topic of this paper. We focus on the characteristics and applications of GLMMs.

Since the appearance of Laird and Ware (1982) linear mixed models have been widely used (e.g. in bio-statistics and environmental statistics) to model longitudinal data. Mixed models extend classical linear regression models by including random or subject-specific effects – next to the (traditional) fixed effects – in the structure for the mean. For distributions from the exponential family, GLMMs extend GLMs by including random effects in the linear predictor. The random effects not only determine the structure of correlation between observations on the same subject, they also take account of heterogeneity among subjects, due to unobserved characteristics.

In an actuarial context Frees et al. (1999, 2001) provide an excellent introduction to linear mixed models and their applications in ratemaking. We will revisit some of their illustrations in the framework of generalized linear mixed models. Using likelihood-based hierarchical generalized linear models, Nelder and Verrall (1997) give an interpretation of traditional credibility models in the framework of GLMs. Hierarchical generalized linear models are GLMMs with random effects that are not necessarily normally distributed; an assumption that is traditionally made. Since the statistical expertise concerning GLMMs is more extensive, this paper puts focus on these models. In addition to traditional credibility models, various other applications are considered as well.

Because both are valuable, estimation and inference in a likelihood-based as well as a Bayesian framework are discussed. In a commercial software package like SAS,[1] the results of a likelihood-based analysis are easy to obtain with standard statistical procedures. Our Bayesian implementation relies on Markov Chain Monte Carlo (MCMC) simulations. The results of the likelihood-based analysis can be used for instance to choose starting values for the chains and to check the reasonableness of the results. In an actuarial context, an important advantage of the Bayesian approach is that it yields the posterior predictive distribution of quantities of interest.

Spatial data and generalized additive mixed models (GAMMs) are outside the scope of this paper. Recent work by Denuit and Lang (2004) and Fahrmeir et al. (2003) considers a Bayesian implementation of a generalized additive model (GAM) for insurance data with a spatial structure.

The paper is organized as follows. Section 2 introduces two motivating data sets which will be analyzed later on. In Section 3 we first recall (briefly) the basic concepts of GLMs and linear mixed models. Afterwards GLMMs are introduced and both maximum likelihood (i.e. pseudo-likelihood or penalized quasi-likelihood and (adaptive) Gauss–Hermite quadrature) and Bayesian inference are discussed. In Section 4 we start with the formulation of basic credibility models as particular GLMMs. The crossed classification model of Dannenburg et al. (1996) is illustrated on a data set. Afterwards, illustrations on workers' compensation insurance data are fully explained. Other interesting applications of GLMMs, for instance in credit risk modelling, are briefly sketched. Finally, Section 5 concludes.

2. Motivating actuarial examples

Two data sets from workers' compensation insurance are considered. With the introduction of these data we want to motivate the need for an extension of GLMs that is appropriate to model correlated – here: longitudinal – data.

2.1. Workers' compensation insurance: Frequencies

The data are taken from Klugman (1992). Here 133 occupation or risk classes are followed over a period of 7 years. Frequency counts in workers' compensation insurance are observed on a yearly basis. Let Count denote the response variable of interest. Possible explanatory variables are Year and Payroll, a measure of exposure denoting scaled payroll totals adjusted for inflation. Klugman (1992) and later on also Scollnik (1996) and Makov et al. (1996) have analyzed these data in a Bayesian context (with no explicit formulation as a GLMM). Exploratory plots for the raw data (not adjusted for exposure) are given in Figs. 1 and 2. A histogram of the complete data set and boxplots of

[1] SAS is a commercial software package, see http://www.sas.com.