

Risk Classification In Non-Life Insurance

Katrien Antonio^{*†} Jan Beirlant[‡]

November 28, 2006

Abstract

Within the actuarial profession a major challenge can be found in the construction of a fair tariff structure. We discuss both a priori and a posteriori rating systems in light of the statistical techniques that are involved. The article introduces basic concepts, illustrates them with real-life actuarial data and summarizes references to complementary literature. Examples of likelihood-based as well as Bayesian estimation are included.

Keywords: non-life insurance, risk classification, generalized linear models, longitudinal data, mixed models, prediction, Bayesian statistics, heavy-tailed regression.

1 Introduction

Within the actuarial profession a major challenge can be found in the construction of a fair tariff structure. In light of the heterogeneity within, for instance, a car insurance portfolio, an insurance company should not apply the same premium for all insured risks. Otherwise the so-called concept of adverse selection will undermine the solvability of the company. ‘Good’ risks, with low risk profiles, pay too much and leave the company, whereas ‘bad’ risks are attracted by the (for them) favorable tariff. The idea behind risk classification is to split an insurance portfolio into classes that consist of risks with a similar profile and to design a fair tariff for each of them. Classification variables typically used in motor third party liability insurance are the age and gender of the policyholder and the type and use of their car.

Being able to identify important risk factors is an important skill for the non-life actuary. When these explanatory variables contain a priori correctly measurable information about the policyholder (or for instance the vehicle or the insured building), the system

^{*}{Katrien.Antonio, Jan.Beirlant}@{econ,wis}.kuleuven.be

[†]Statistics Section and AFI Research Center, KU Leuven, W. de Croylaan 54, 3001 Heverlee, Belgium.

[‡]Full professor, Statistics Section, KU Leuven, W. de Croylaan 54, 3001 Heverlee, Belgium.

is called an a priori classification scheme. However, an a priori system will not be able to identify all important factors because some of them can not be measured or observed. Think for instance of aggressiveness behind the wheel or the swiftness of reflexes. Thus, despite of the a priori rating system, tarification cells will not be completely homogeneous. For that reason, an a posteriori rating system will re-evaluate the premium by taking the history of claims of the insured into account.

Due to the quantitative nature of both a priori and a posteriori rating, one of the primary attributes of an actuary should be the successful application of up-to-date statistical techniques in the analysis of insurance data. Therefore, this article highlights current techniques involved in this area of actuarial statistics. The article introduces basic concepts, illustrates them with real-life actuarial data and summarizes references to complementary literature. Examples of likelihood-based as well as Bayesian estimation are included where the latter has the advantage that it provides the analyst with the full predictive distribution of quantities of interest.

Remark 1 *Link With Statistical Techniques For Loss Reserving* *The statistical techniques discussed here in the context of risk classification, also provide a useful framework for a stochastic approach of the loss reserving problem in actuarial science. In this context the data are displayed in a traditional run-off triangle or variations of it. See [1] and [2] for connections with reserving techniques.*

2 Regression Models For A Priori Risk Classification

In order to build a tariff that reflects the various risk profiles in a portfolio in a reasonable way, actuaries will rely on regression techniques. Typical response variables involved in this process are the number of claims (or the claims frequency) on the one hand and its corresponding severity (i.e. the amount the insurer will have to pay, given that a claim occurred) on the other hand.

2.1 Generalized Linear Models

The history of generalized linear models (GLMs) in actuarial statistics goes back to the actuarial illustrations in the standard text by [3]. See [4] for an overview in actuarial science. GLMs extend the framework of general (normal) linear models to the class of distributions from the exponential family. A whole variety of possible outcome measures (like counts, binary and skewed data) can be modelled within this framework. This paper uses the canonical form specification of densities from the exponential family, namely

$$f(y) = \exp\left(\frac{y\theta - \psi(\theta)}{\phi} + c(y, \phi)\right) \quad (1)$$

where $\psi(\cdot)$ and $c(\cdot)$ are known functions, θ is the natural and ϕ the scale parameter. Members of this family, often used in actuarial science, are the normal, the Poisson, the binomial and the gamma distribution. Instead of a transformed data vector, GLMs model a transformation of the mean as a linear function of explanatory variables. In this way

$$g(\mu_i) = \eta_i = (\mathbf{X}\boldsymbol{\beta})_i \quad (2)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ contains the model parameters and \mathbf{X} ($n \times p$) is the design matrix. g is the link function and η_i is the i^{th} element of the so-called linear predictor. In a likelihood-based approach, the unknown but fixed regression parameters in $\boldsymbol{\beta}$ are estimated by solving the maximum likelihood equations with an iterative numerical technique (such as Newton-Raphson). In a Bayesian approach, priors are assigned to every parameter in the model specification and inference is based on samples generated from the corresponding full posterior distributions.

Illustration 1 *Poisson Regression For Claims Frequencies* For detailed case-studies on Poisson regression for claim counts, [5] and [6] contain nice examples.

2.2 Flexible, Parametric Families of Distributions And Regression

Modelling the severity of claims as a function of their risk characteristics (given as covariate information) might require statistical distributions outside the exponential family. Distributions with a heavy tail, for instance. Principles of regression within such a family of distributions are illustrated here with the Burr XII and the GB2 ('generalized beta of the second kind') distribution. More details on Burr regression are in [7] and for GB2 regression [8] and [9] are useful.

Illustration 2 *Fire Insurance Portfolio* The cumulative distribution function for the Burr Type XII and the GB2 distribution are given by

$$F_{Burr,Y}(y) = 1 - \left(\frac{\beta}{\beta + y^\tau} \right)^\lambda, \quad y > 0, \beta, \lambda, \tau > 0, \quad (3)$$

$$F_{GB2,Y}(y) = B \left(\frac{(y/b)^a}{1 + (y/b)^a}; \gamma_1, \gamma_2 \right), \quad y > 0, a \neq 0, b, \gamma_1, \gamma_2 > 0, \quad (4)$$

where $B(\cdot, \cdot)$ is the incomplete Beta function. Say, the available covariate information is in \mathbf{x} ($1 \times p$). By allowing one or more of the parameters in (3) or (4) to vary with \mathbf{x} , a Burr or GB2 regression model is built. To illustrate this approach, consider a fire insurance portfolio (see [7]) which consists of 1,823 observations. We want to assess how the loss distribution changes with the sum insured and the type of building. Claims expressed as a fraction of the sum insured are used as the response. Explanatory variables

are the type of building and the sum insured. Parameters are estimated with maximum likelihood. Residual QQplots like those in Figure 1 can be used to judge the goodness-of-fit of the proposed regression models. [7] explains their construction.

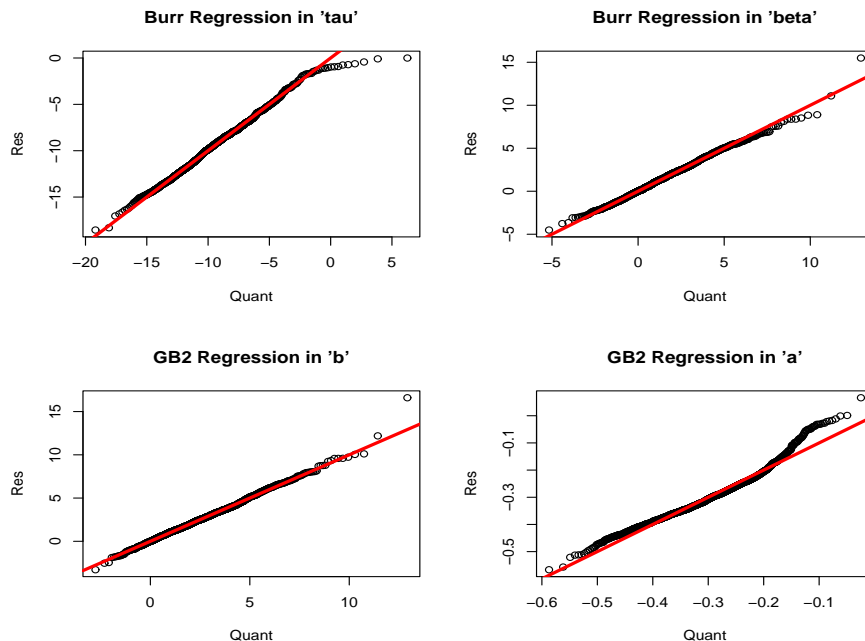


Figure 1: *Fire Insurance Portfolio: residual QQplots for Burr and GB2 regression.*

3 A Posteriori Ratemaking

To update the a priori tariff or to predict future claims when historical claims are available over a period of insurance, actuaries can use statistical models for longitudinal or panel data (i.e. data observed on a group of subjects, over time). These statistical models (also known as ‘mixed models’ or ‘random-effects models’) generalize the so-called credibility models from actuarial science. Bonus-malus systems for car insurance policies are another example of a posteriori corrections to a priori tariffs: insured drivers reporting a claim to the company will get a malus, causing an increase of their insurance premium in the next year. More details are e.g. in [6].

The credibility ratemaking problem is concerned with the determination of a risk premium that combines the observed, individual claims experience of a risk and the experience regarding related risks. The framework of our discussion of this a posteriori rating scheme is the concept of generalized linear mixed models (GLMMs). For a historical, analytical discussion of credibility, [10] is a good reference. [11]-[15] will provide additional background. [16]-[17] and [18] discuss explicitly the connection between actuarial credi-

bility schemes and (generalized) linear mixed models and contain more detailed examples. [19] yet provides another statistical approach using copulas instead of random effects.

GLMMs extend GLMs by allowing for random, or subject-specific, effects in the linear predictor. Say we have a data set at hand consisting of N policyholders. For each subject i ($1 \leq i \leq N$) n_i observations are available. These are the claims histories. Given the vector \mathbf{b}_i with the random effects for subject (or cluster) i , the repeated measurements Y_{i1}, \dots, Y_{in_i} are assumed to be independent with a density from the exponential family

$$f(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) = \exp\left(\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right), \quad j = 1, \dots, n_i. \quad (5)$$

The following (conditional) relations then hold

$$\mu_{ij} = E[Y_{ij}|\mathbf{b}_i] = \psi'(\theta_{ij}) \quad \text{and} \quad \text{Var}[Y_{ij}|\mathbf{b}_i] = \phi\psi''(\theta_{ij}) = \phi V(\mu_{ij}) \quad (6)$$

where $g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i$. $g(\cdot)$ is called the link and $V(\cdot)$ the variance function. $\boldsymbol{\beta}$ ($p \times 1$) denotes the fixed effects parameter vector and \mathbf{b}_i ($q \times 1$) the random effects vector. \mathbf{x}_{ij} ($p \times 1$) and \mathbf{z}_{ij} ($q \times 1$) contain subject i 's covariate information for the fixed and random effects, respectively. The specification of the GLMM is completed by assuming that the random effects, \mathbf{b}_i ($i = 1, \dots, N$), are mutually independent and identically distributed with density function $f(\mathbf{b}_i|\boldsymbol{\alpha})$. Hereby $\boldsymbol{\alpha}$ denotes the unknown parameters in the density. Traditionally, one works under the assumption of (multivariate) normally distributed random effects with zero mean and covariance matrix determined by $\boldsymbol{\alpha}$. The random effects \mathbf{b}_i represent unobservable, individual characteristics of the policyholder. Correlation between observations on the same subject arises because they share the same random effects. [18] provides a discussion of likelihood-based and Bayesian estimation for GLMMs; with references to the literature and worked-out examples.

Illustration 3 Workers' Compensation Insurance *The data are taken from [20]. 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. Possible explanatory variables are Year and Payroll, a measure of exposure denoting scaled payroll totals adjusted for inflation. The following models are considered*

$$Y_{ij}|\mathbf{b}_i \sim \text{Poisson}(\mu_{ij})$$

$$\text{where } \log(\mu_{ij}) = \log(\text{Payroll}_{ij}) + \beta_0 + \beta_1 \text{Year}_{ij} + b_{i,0} \quad (7)$$

$$\text{versus } \log(\mu_{ij}) = \log(\text{Payroll}_{ij}) + \beta_0 + \beta_1 \text{Year}_{ij} + b_{i,0} + b_{i,1} \text{Year}_{ij}. \quad (8)$$

Hereby Y_{ij} represents the j^{th} measurement on the i^{th} subject of the response **Count**. β_0 and β_1 are fixed effects and $b_{i,0}$, versus $b_{i,1}$, is a risk class specific intercept, versus slope. It is assumed that $\mathbf{b}_i = (b_{i,0}, b_{i,1})' \sim N(\mathbf{0}, \mathbf{D})$ and that, across subjects, random effects are independent. The results of both a maximum likelihood (Penalized Quasi-Likelihood

and adaptive Gauss-Hermite quadrature) and a Bayesian analysis are given in Table 1. The models were fitted to the data set without the observed Count_{i7} , to enable out-of-sample prediction later on. To illustrate prediction with model (8), Table 2 compares the predictions for some selected risk classes with the observed values. Predictive distributions obtained with a Bayesian analysis are illustrated in Figure 2.

	PQL		adaptive G-H		Bayesian	
	Est.	SE	Est.	SE	Mean	90% Cred. Int.
Model (7)						
β_0	-3.529	0.083	-3.557	0.084	-3.565	(-3.704, -3.428)
β_1	0.01	0.005	0.01	0.005	0.01	(0.001, 0.018)
δ_0	0.790	0.110	0.807	0.114	0.825	(0.648, 1.034)
Model (8)						
β_0	-3.532	0.083	-3.565	0.084	-3.585	(-3.726, -3.445)
β_1	0.009	0.011	0.009	0.011	0.008	(-0.02, 0.04)
δ_0	0.790	0.111	0.810	0.115	0.834	(0.658, 1.047)
δ_1	0.006	0.002	0.006	0.002	0.024	(0.018, 0.032)
$\delta_{0,1}$	/	/	0.001	0.01	0.006	(-0.021, 0.034)

Table 1: *Workers' compensation data (frequencies): results of maximum likelihood and Bayesian analysis.* $\delta_0 = \text{Var}(b_{i,0})$, $\delta_1 = \text{Var}(b_{i,1})$ and $\delta_{0,1} = \delta_{1,0} = \text{Cov}(b_{i,0}, b_{i,1})$

Class	Actual Values		Expected number of claims					
	Payroll _{i7}	Count _{i7}	PQL		adaptive G-H		Bayesian	
			Mean	s.e.	Mean	s.e.	Mean	90% Cred. Int.
11	230	8	11.294	2.726	11.296	2.728	12.18	(5,21)
20	1,315	22	33.386	4.109	33.396	4.121	32.63	(22,45)
70	54.81	0	0.373	0.23	0.361	0.23	0.416	(0,2)
89	79.63	40	47.558	5.903	47.628	6.023	50.18	(35,67)
112	18,810	45	33.278	4.842	33.191	4.931	32.66	(21,46)

Table 2: *Workers' compensation data (frequencies): predictions for selected risk classes.*

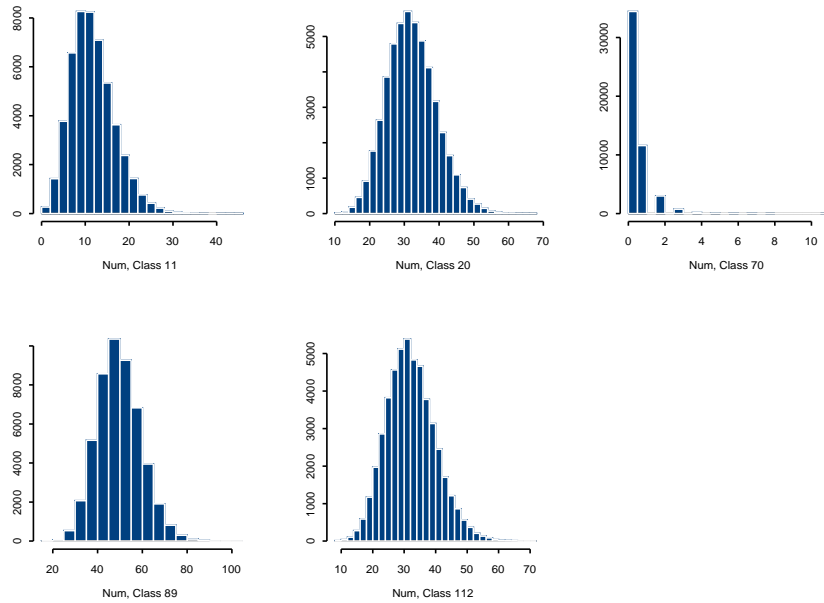


Figure 2: *Workers Compensation Insurance (Counts): predictive distributions for selection of risk classes.*

4 Zero-Inflated, Additive And Spatial Regression Models For A Priori And A Posteriori Ratemaking

This Section contains references to some more advanced regression techniques for both cross-sectional data (i.e. one observation per subject, as in Section 2) and panel data.

Dealing with regression models for claim counts, the huge number of zeros (i.e. no claim events) is often apparent. For data sets where the inflated number of zeros causes a bad fit of the regular Poisson or negative binomial distribution, zero-inflated regression models ([21]) provide an alternative. See [22] for a discussion of zero-inflated regression models for a priori classification schemes for count data.

In the modelling of severity data that consist of exact zeros and strictly positive payments, so-called two-part regression models are often used. They specify separate regression models for $Y = 0$ and $Y > 0$. See [2] for an illustration.

So far, only regression models with a linear structure for the mean or a transformation of the mean have been discussed. To allow for more flexible relationships between a response and a covariate, generalized additive models (GAMs) are available, see e.g. [2] for several actuarial examples. When the place of residence of a policyholder is available as covariate information, spatial regression models allow to bring this into account. See [23] for an illustration with spatial GAMs for cross-sectional observations on claim counts and severities.

Evidently, the mixed models for a posteriori ratemaking discussed above can be extended to zero-inflated, generalized additive mixed models (GAMMs) and spatial models to allow for more flexibility in model building. See e.g. [2] for an example with GAMMs.

References

- [1] Antonio, K., Beirlant, J., Hoedemakers, T. & Verlaak, R. (2006). Lognormal mixed models for reported claims reserving, *North American Actuarial Journal* **10(1)**, 30-48.
- [2] Antonio, K. & Beirlant, J. (2006). Issues in claims reserving and credibility: a semiparametric Approach with Mixed Models. Working Paper, available online at www.econ.kuleuven.be/insurance.
- [3] McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*. Monographs on statistics and applied probability, Chapman and Hall, New York.
- [4] Haberman, S. & Renshaw, A.E. (1996). Generalized linear models and actuarial science, *The Statistician* **45(4)**, 407-436.
- [5] Denuit, M. & Charpentier, A. (2005). *Mathématiques de l'assurance non-vie: tarification et provisionnement (Tome 2)*. Economica, Paris.
- [6] Denuit, M., Marechal, X., Pitrebois, S. & Walhin, J.-F. (2007). *Actuarial modelling of claim counts: risk classification, credibility and bonus-malus scales*. Wiley.
- [7] Beirlant, J., Goegebeur, Y., Verlaak, R. & Vynckier, P. (1998). Burr regression and portfolio segmentation, *Insurance: Mathematics and Economics*, **23**, 231-250.
- [8] Cummins, D.J., Dionnes, G., McDonald, J.B. & Pritchett, M.B. (1990). Applications of the GB2 family of distributions in modelling insurance loss processes, *Insurance: Mathematics and Economics*, **9**, 257-272.
- [9] Sun, J., Frees, E.W. & Rosenberg, M.A. (2006). Heavy-tailed longitudinal data modelling using copulas. Working Paper, available online at <http://research.bus.wisc.edu/jfrees/>.
- [10] Dannenburg, D.R., Kaas, R. & Goovaerts, M.J. (1996). *Practical actuarial credibility models*. Institute of actuarial science and econometrics, University of Amsterdam, Amsterdam.
- [11] Dionne, G. & Vanasse, C. (1989). A generalization of actuarial automobile insurance rating models: the negative binomial distribution with a regression component, *ASTIN Bulletin*, **19**, 199-212.

- [12] Pinquet, J. (1997). Allowance for cost of claims in bonus-malus systems, *ASTIN Bulletin*, **27(1)**, 33-57.
- [13] Pinquet, J. (1998). Designing optimal bonus-malus systems from different types of claims, *ASTIN Bulletin*, **28(2)**, 205-229.
- [14] Pinquet, J., Guillén, M. & Bolancé, C. (2001). Allowance for age of claims in bonus-malus systems, *ASTIN Bulletin*, **31(2)**, 337-348.
- [15] Bolancé, C., Guillén, M. & Pinquet, J. (2003). Time-varying credibility for frequency risk models: estimation and tests for autoregressive specifications on random effects, *Insurance: Mathematics and Economics*, **33**, 273-282.
- [16] Frees, E.W., Young, V.R. & Luo, Y. (1999). A longitudinal data analysis interpretation of credibility models, *Insurance: Mathematics and Economics*, **24(3)**, 229-247.
- [17] Frees, E.W., Young, V.R. & Luo Y. (2001). Case studies using panel data models, *North American Actuarial Journal*, **5(4)**, 24-42.
- [18] Antonio, K. & Beirlant, J. (2007). Applications of Generalized Linear Mixed Models in Actuarial Statistics, *Insurance: Mathematics and Economics*, **40**, 58-76.
- [19] Frees, E.W. & Wang, P. (2005). Credibility using copulas, *North American Actuarial Journal*, **9(2)**, 31-48.
- [20] Klugman, S. (1992). *Bayesian statistics in actuarial science with emphasis on credibility*. Kluwer, Boston.
- [21] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1-14.
- [22] Boucher, J.-P., Denuit, M. & Guillén, M. (2005). Risk classification for claim counts: mixed Poisson, zero-inflated mixed Poisson and hurdle models. Working Paper, available online at www.actu.ucl.ac.be.
- [23] Denuit, M. & Lang, S. (2004). Non-life ratemaking with Bayesian GAMs, *Insurance: Mathematics and Economics*, **35(3)**, 627-647.