PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 106
Ubezpieczenia emerytalne, społeczne i metody aktuarialne 2010

Viera Pacáková, Hana Boháčová
University of Pardubice

PREMIUM CALCULATION IN A HETEROGENEOUS PORTFOLIO OF POLICIES[*]

Summary: The calculation of fair premiums is the main actuarial problem in insurance business. We assume a heterogeneous portfolio of policies such as motor insurance portfolio. The concept of a mixture distribution will be convenient to use and show that the Poisson model for the number of claims and the gamma distribution for the modelling insurance losses are useful in such a case.

For a high-volume class of business such as private motor insurance, which has well established rating factors, one of the most common pricing techniques used is generalized linear modelling. This paper deals with technical aspects of this approach and presents its application for estimating the net premiums according to the rating factors across a group of policyholders.

Key words: mathematical modeling claim frequency & claim anount, heterogeneous portfolio of policies, non-life insurance.

1. Mathematical modeling in non-life insurance

Mathematical modeling is an important tool used in the non-life insurance, it is a part of the risk theory. A simple mathematical model with a few parameters often describes a large heterogeneous insurance portfolio quite well. The two quantities – claim frequency and claim amount – are especially modeled.

The claim frequency indicates the number of claims related to one insurance policy of a given type during one year. It can go up in non-negative integer values and therefore it is modeled by the discrete probability distributions – binomial, Poisson, negative binomial or mixed Poisson distribution.

The claim amount models proceed from a probability distribution of a variable indicating an individual claim amount per one average. They use the continuous probability distributions – lognormal, exponential, gamma, Weibull or Pareto distribution. In the heterogeneous portfolio the claim frequency is most often modeled by the negative-binomial distribution and the claim amount by the Pareto distribution, for more details see [Pacáková 2004].

[*] This paper originated within the framework of the research grant of the Czech Science Foundation GAČR – GA 491866.

As the claim frequency is a discrete random variable and the claim amount is a non-negative one they cannot be modeled by the normal linear regression models. The generalized linear models (GLMs) are suited to the non-normal data analysis and therefore they are applicable to the modeling of claim frequency and claim amount.

2. Generalized linear models

The generalized linear models are an adaptable generalization of often used linear regression models. They enable an efficient approach to many practical situations among others in the actuarial science even if the linear regression model presumptions are not satisfied.

Suppose that Y_i, $i = 1,...,n$ are independent identically distributed random variables with distribution from the exponential family of distributions, y_i, $i = 1,...,n$ are the observed values of Y_i. The exponential family of distribution [Anderson et al. 2005] contains distributions whose density function (or probability mass function in a discrete case) can be written down in the following form:

$$f\left(y_i, \vartheta_i, \varphi\right) = \exp\left\{\frac{a\left(y_i\right)b\left(\vartheta_i\right) - c\left(\vartheta_i\right)}{h\left(\varphi\right)} + d\left(y_i, \varphi\right)\right\}, \tag{1}$$

where functions $a(y_i)$, $b(\vartheta_i)$, $c(\vartheta_i)$, d(y_i, φ) and $h(\varphi)$ are specified in advance, ϑ_i are parameters related to the mean and ϕ is a scale parameter related to the variance. Other requirements imposed on these functions are as follows:

- $h(\varphi)$ is positive and continuous,
- $c(\vartheta)$ is a twice differentiable convex function.

If $a\left(y_i\right)$ is the identity function then the distribution is said to be in the canonical form. If in addition $b\left(\vartheta\right)$ is the identity function as well and the scale parameter ϕ is known then ϑ is called the canonical parameter.

The exponential family of distributions is quite large. Normal, Poisson, gamma or binomial distributions are probably the most common members of this family. The distribution is fully specified by its mean and variance. Let us denote

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}. \tag{2}$$

Let us suppose the mean of Y_i is

$$\mathrm{E}(Y_i) = \mu_i = g^{-1}\left[\left(X\beta\right)_i\right], \tag{3}$$

the link function *g* is a differentiable and monotonic function of components of the linear predictor *Xβ* The design matrix *X* is of dimension $n \times k$ and is known. The *k*-dimensional unknown vector parameter *β* needs to be estimated by some suitably chosen method.

There are many commonly used link functions and the choice can be somewhat arbitrary. When using a distribution function with a canonical parameter ϑ such a link function can be found which allows for *X'**Y*** to be a sufficient statistic for *β*. It is called a canonical link function. For more details on the canonical link function see [Anderson 2005] or [Kaas 2001].

When we look at the linear regression model we can see the link function is an identity one in this case – $\mu = X\beta$, the normal probability distribution and constant variance presumption is necessary for the estimator of *β*. Why is this not always enough? One reason can be that in some cases the values of y_i are naturally required to be positive. It often happens among others when solving some actuarial problems, for instance the premium amount, that expected cost of claims or claim number may not be negative. The non-negativity assumption disables normality and it raises doubts that the variance of *Y* tends to zero as the mean of *Y* tends to zero. That denies the constant variance so it seems to be reasonable to expect the variance to be a function of the mean, as generalized linear models do:

$$Var(Y_i) = \frac{\varphi V(\mu_i)}{\omega_i}. \tag{4}$$

$V(\bullet)$ is known variance function and ω_i is a constant assigning a weight or credibility to the *i*-th observation.

The main task when fitting a GLM is to estimate the parameter β. The maximum likelihood method is usually used for this purpose and some iterative procedure is needed get the estimate. (e.g. Fisher scoring method, for more details see [de Jong, Heller 2008]).

3. Application

We assume a heterogeneous portfolio of 500 policies such as motor hull insurance portfolio. We know the claim amount and number of claims during one year for each policy and we have information on gender and residence (big city, small twn or country) of each policyholder. We would like to determine the net premium amount on the basis of sex and residence of the policyholder. There are 50 policies with at least one claim during the reference year and 450 policies with no claims in our portfolio. The structure of gender and residence of the owners of the remaining 450 policies with no claims is in Table 1.

Table 1. Structure of gender and residence of the policyholders of policies with no claims

Gender\Residence	Big city	Small town	Country
Male	161	40	71
Female	142	12	25

Source: [Anderson et al. 2005].

Data on the policies with claims are in Table 2. One in columns: Male, Female, Big city, Country and Small town indicates that the value in the column heading is true for the corresponding policy, zero indicates the value is not true. The Claim amount column contains the total amount of all the claims connected to that policy during the year under consideration.

The aim is to find a suitable generalized linear model for the claim amount and another one for the claim frequency. The net premium amount will subsequently be determined as a product of appropriate expected value of claim amount and expected value of claim frequency.

4. GLM for the claim amount

The input values for the first model are values from the column Claim amount in Table 2 divided by the values from the Claim frequency column in the same table. The design matrix X contains three columns of Table 2: Male, Big city and Country. The remaining columns: Female and Small town are not involved as they are dependent upon the other columns – column Female depends on the Male column (the sum of the values in these two columns is always one) and Small town depends on Big city and Country columns in the same way.

Table 2. Claim amount and claim frequency

Observation	Male	Female	Big city	Country	Small town	Claim amount	Claim frequency
1	2	3	4	5	6	7	8
1	1	0	0	0	1	1.117514	1
2	1	0	0	1	0	1.925891	1
3	0	1	1	0	0	9.960349	1
4	0	1	1	0	0	52.76903	1
5	0	1	1	0	0	34.67459	1
6	0	1	0	1	0	10.99608	1
7	1	0	1	0	0	770.7137	1
8	1	0	1	0	0	7.413328	1
9	1	0	1	0	0	961.1342	1
10	0	1	0	1	0	0.128025	1

1	2	3	4	5	6	7	8
11	0	1	0	0	1	2.721808	1
12	0	1	0	1	0	4.037756	1
13	0	1	0	0	1	38.58484	1
14	0	1	1	0	0	10.94166	1
15	1	0	0	1	0	60.73693	1
16	1	0	0	1	0	8.249735	1
17	1	0	0	1	0	1.99354	1
18	0	1	1	0	0	2.307934	1
19	1	0	0	0	1	78.50715	1
20	0	1	0	1	0	0.289212	1
21	0	1	1	0	0	87.88076	1
22	0	1	1	0	0	60.18832	1
23	1	0	0	0	1	187.4077	1
24	1	0	1	0	0	918.6962	1
25	1	0	1	0	0	4.683388	1
26	1	0	0	1	0	0.120435	1
27	1	0	0	1	0	3.025471	1
28	0	1	0	0	1	27.74861	1
29	0	1	1	0	0	162.5094	1
30	0	1	1	0	0	353.7966	1
31	1	0	1	0	0	0.235596	1
32	1	0	1	0	0	154.2416	1
33	1	0	1	0	0	109.6394	1
34	0	1	1	0	0	62.00207	1
35	1	0	1	0	0	42.33152	1
36	1	0	1	0	0	25.94723	1
37	1	0	1	0	0	395.5816	1
38	0	1	1	0	0	33.44059	1
39	1	0	1	0	0	311.4383	1
40	1	0	0	1	0	11.72739	1
41	1	0	0	1	0	24.40291	1
42	0	1	0	0	1	42.97137	1
43	0	1	1	0	0	549.8948	2
44	0	1	1	0	0	119.2653	2
45	1	0	1	0	0	121.8874	2
46	1	0	1	0	0	568.9089	2
47	1	0	1	0	0	290.9324	2
48	0	1	1	0	0	135.1506	2
49	1	0	1	0	0	87.5612	2
50	1	0	1	0	0	130.065	2

Source: [Anderson et al. 2005].

The GLM fit was done in R software. According to the residual deviance the gamma distribution and inverse link were selected. The estimate of the parameters β from equation (3) is

$$\hat{\beta} = (0.022726, -0.007952, -0.010903, 0.076739)' . \tag{5}$$

The dispersion parameter ϕ=1.684 was counted and 7 Fisher scoring iterations were done.

The first element of $\hat{\beta}$ is the intercept, the second one is the multiplier of the Male value, the third one corresponds to the Big city column and the last one to Country. The expected values of the claim amount for particular groups of policy holders are counted according to the equation (3), the link function g is the inverse one:

$$g = \frac{1}{\mu}. \tag{6}$$

The resulting expected values are to be found in Table 3.

Table 3. Expected values of the individual claim amount

Gender& Residence	Claim amount expected value
Male, Big city	258.331
Male, Small town	67.686
Male, Country	10.927
Female, Big city	84.581
Female, Small town	44.002
Female, Country	10.054

Source: own calculation.

5. GLM for the claim frequency

A similar procedure was done for the claim frequency. The design matrix **X** had the same columns as in the case of the claim amount fitting. The observation vector contains the claim frequencies from Table 2 together with the 450 zeros representing the policies with no claims in the year under consideration. The Poisson distribution together with logarithmic link function $g = \ln \mu$ gave the best fit.

The estimate of the parameters is at the time in the form:

$$\hat{\beta} = (-2.04666, -0.11852, -0.03668, -0.06527)' . \tag{7}$$

The dispersion parameter was taken to be 1 (as it is standard in case of the Poisson distribution). Six Fisher scoring iterations were needed to reach the estimate. Table 4 that follows contains the estimates of the claim frequencies for particular groups of the policyholders.

Table 4. Expected values of the claim frequency

Gender& Residence	Claim frequency expected value
Male, Big city	0.111
Male, Small town	0.115
Male, Country	0.107
Female, Big city	0.125
Female, Small town	0.129
Female, Country	0.121

Source: own calculation.

6. Net premium calculation

The proposed net premium for each group of policyholders can now be counted as a product of the expected value of individual claim and the expected value of the claim frequency for that group.

Table 5. Calculated net premium

Gender& Residence	Premium
Male, Big city	28.571
Male, Small town	7.766
Male, Country	1.174
Female, Big city	10.531
Female, Small town	5.684
Female, Country	1.217

Source: own calculation.

Literature

Anderson D. et al., *A Practitioner's Guide to Generalized Linear Models – A Foundation for Theory, Interpretation and Application*, Watson Wyatt Worldwide, London 2005.

Cipra T., *Pojistná matematika – teorie a praxe*, Ekopress, Praha 2006.

de Jong P., Heller G.Z., *Generalized Linear Models for Insurance Data*, Cambridge University Press, New York 2008.

Kaas R. et al., *Modern Actuarial Risk Theory*, Kluwer Academic Publishers, Boston – Dordrecht – London 2001.

Pacáková V., *Aplikovaná poistná štatistika* (in Slovak), Iura Edition, Bratislava 2004.

Šoltés E., *Regresná a korelačná analýza s aplikáciami* (in Slovak), Iura Edition, Bratislava 2008.

KALKULACJA SKŁADKI W HETEROGENICZNYCH PORTFELACH POLIS

Streszczenie: Kalkulacja składki ubezpieczeniowej jest ważnym zagadnieniem związanym z matematyką aktuarialną oraz samą działalnością ubezpieczeniową. Artykuł prezentuje zastosowanie metody opartej na konstrukcji portfeli heterogenicznych w szacowaniu składki dla ubezpieczeń komunikacyjnych. Koncepcja wykorzystywania mieszanki rozkładów jest wygodna w stosowaniu i pokazuje, że rozkład Poissona jest odpowiedni do modelowania liczby szkód, rozkład gamma zaś – do modelowania (wysokości) strat ubezpieczeniowych.

Do klasyfikacji ubezpieczeń komunikacyjnych (a także innych *high-volume*) wykorzystuje się uogólnione modelowanie liniowe. Prezentowany artykuł dotyczy technicznych aspektów stosowania portfeli heterogenicznych w kalkulacji składki ubezpieczeniowej netto w zależności od czynników dotyczących całego portfela ubezpieczeń.