PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 60 2009
Inwestycje finansowe i ubezpieczenia – tendencje światowe a polski rynek

Hana Boháčová, Bohdan Linda
University of Pardubice

GENERALIZED LINEAR MODELS IN ACTUARIAL SCIENCE

The generalized linear models (GLMs) 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 2005] contains distributions whose density function (or probability mass function in a discrete case) can be written down in the following form:

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

where functions $a(y_i)$, $b(\vartheta_i)$, $c(\vartheta_i)$, $d(y_i,\phi)$ and $h(\phi)$ 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:

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

If $a(y_i)$ is the identity function then the distribution is said to be in the canonical form. If in addition $b(\vartheta)$ 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

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

Let us suppose the mean of Y_i is

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

the link function g is differentiable and monotonic function of components of the linear predictor $X\beta$. 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 link function can be found which allows for $X'\boldsymbol{Y}$ 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, this happens often among others when solving some of the actuarial problems, for instance the premium amount, 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:

$$Var\left(Y_i\right)=\frac{\phi V\left(\mu_i\right)}{\omega_i}.$$

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

Let us try to explain one of the actuarial applications of GLMs. The claims originating in a particular year often cannot be finalized in the same year. However they should be connected to the year for which the premium was paid actually. This means that reserves have to be kept due to claims which are known to exist but for which the appropriate size is not known yet. We can use a so called run-off triangle [Kaas 2001] which is created in the following way – we start in the year *1* with a given number of insurance policies and known premium amount paid in that year (all the claims occur-

ring in the year *1* have to be paid from it). We know some claims from *1* have been paid in the same year, some of them later on – in years *2*, *3*, …, in year *k* at the latest. Further for the year *2* we know the total payments connected to the claims arisen in this year and settled in years *2*, *3*, …., *k*. And so on in a similar way, the last known entry is the total paid amount concerning the premiums from year *k* and settled in the same year. Let us denote the total payments originated in the *i*-th year and settled in the year *i+j-1* with $x_{i,j}$ for $i = 1, 2, \ldots, k$ and $j=1, \ldots, k-i+1$. All the observed total payments can be written in a table as bellow:

Table 1.

		Development year						
		1	2	3	…	$k-2$	$k-1$	k
Year of origin	1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	…	$x_{1,k-2}$	$x_{1,k-1}$	$x_{1,k}$
	2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	…	$x_{2,k-2}$	$x_{2,k-1}$	
	3	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	…	$x_{3,k-2}$		
	⋮	⋮	⋮	⋮	⋱			
	$k-2$	$x_{k-2,1}$	$x_{k-2,2}$	$x_{k-2,3}$				
	$k-1$	$x_{k-1,1}$	$x_{k-2,2}$					
	k	$x_{k,1}$						

Source: own calculations.

The sums of values in the up-right bottom-left diagonals with $i + j - 1 = c$ give the total payments made in the year x_c for $c = 1, \ldots, k$. What we have to do now is to predict the future values that are missing in the bottom-right part of the table above. Such a prediction enables to estimate the necessary reserves for the period *2, …, k*. (The sum of predicted values in each of the diagonals (for $c>k$) is the estimate of the total of the claims that will have to be paid in the *c*-th calendar year from the premiums that were collected during years *2, …, k*, which are in fact the estimates of the reserves to be kept.) We can describe this situation with an uncomplicated GLM.

Let us denote $X_{i,j}$ - the random variables denoting the total paid amounts for the year of origin *i* and the year of development *j*, (it means the appropriate claims were paid in the year x_{i+j-1},) $i, j = 1, 2, \ldots, k$. Suppose each of these random variables may be affected by the year of origin with the effect of α_i, the year of development with the effect of β_j and the year when the claims were paid with the effect of γ_c and take into consideration a multiplicative model:

$$X_{i,j} \approx \alpha_i \beta_j \gamma_c, \; i, j, c = 1, 2, \ldots k, \; c = i + j - 1. \tag{1}$$

If we assume that the random variables $X_{i,j}$ are independent and that their distribution belongs to the exponential family, the model (1) can be understood as a generalized linear model with a logarithmic link function:

$$\ln X_{i,j} \approx \ln \alpha_i + \ln \beta_j + \ln \gamma_c .$$

We know the observations stated in the table 1, they are the values of $X_{i,j}$ for $i + j \le k+1$, and we need to estimate the unknown effects α_i, β_j and γ_c based on these values. When we have the estimates $\hat{\alpha}_i$, $\hat{\beta}_j$ and $\hat{\gamma}_c$ we can easily estimate the remaining $X_{i,j}$:

$$\hat{X}_{i,j} = \hat{\alpha}_i \hat{\beta}_j \hat{\gamma}_c , \; i + j > k+1 . \tag{2}$$

The problem can be that we have $3k$ of unknown parameters and $\frac{k+k^2}{2}$ of available observations (in the table 1). To be able to estimate the parameters we have to insist on

$$\frac{k+k^2}{2} \ge 3k,$$

which is satisfied for $k \ge 5$. As we can not assure this in general we have to decrease the number of unknown parameters in some reasonable way. (With the respect to the quality of the estimates it is always better when the number of the observations exceeds the number of the estimated parameters expressively. So even if $k>5$ it is better to decrease the number of the unknown effects.) Sometimes we know from the experience in the given kind of policies that some effects are insignificant. Then we can equate them to 1 and simplify the model according to this. We also can add some extra constraints (that are somehow logically connected to the specific situation we have) for the estimated parameters which increases the number of available equations. There are several methods derived for IBNR model described above. Some of them can be found in [Pacáková 2004] or [Kaas 2001].

An iterative approach can be applied as well. One of its variants is as follows. Let's assume the calendar year effects γ_c are insignificant, so we can use $\gamma_c = 1$, $c = 1, 2, \ldots, k$. When we add the assumption of the exponential distribution we get from (1) following model:

$$X_{i,j} \sim Exponential(\alpha_i \beta_j) , \; i = 1, 2, \ldots, k, \; j = 1, \ldots, k - i + 1. \tag{3}$$

Our task is to estimate α_i and β_j for $i, j = 1, \ldots, k$ from the observed paid claims $x_{i,j}$ in the table 1. The maximum likelihood equations for this case can be simplified to the common form

$$x_{i,j} = \alpha_i \beta_j , \tag{4}$$

which needs to be satisfied for $i, j = 1, 2, \ldots, k$. Thus we can modify the table 1:

Table 2.

		Development year							Total
		1	2	3	…	$k-2$	$k-1$	k	
Year of origin	1	$\alpha_1\beta_1$	$\alpha_1\beta_2$	$\alpha_1\beta_3$	…	$\alpha_1\beta_{k-2}$	$\alpha_1\beta_{k-1}$	$\alpha_1\beta_k$	R_1
	2	$\alpha_2\beta_1$	$\alpha_2\beta_2$	$\alpha_2\beta_3$	…	$\alpha_2\beta_{k-2}$	$\alpha_2\beta_{k-1}$		R_2
	3	$\alpha_3\beta_1$	$\alpha_3\beta_2$	$\alpha_3\beta_3$	…	$\alpha_3\beta_{k-2}$			R_3
	⋮	⋮	⋮	⋮	⋰				⋮
	$k-2$	$\alpha_{k-2}\beta_1$	$\alpha_{k-2}\beta_2$	$\alpha_{k-2}\beta_3$					R_{k-2}
	$k-1$	$\alpha_{k-1}\beta_1$	$\alpha_{k-1}\beta_2$						R_{k-1}
	k	$\alpha_k\beta_1$							R_k
Total		C_1	C_2	C_3	…	C_{k-2}	C_{k-1}	C_k	

Source: own calculations.

We can derive two systems of equations from the table 2. One for its rows:

$$\begin{aligned}
\alpha_1\left(\beta_1+\beta_2+\beta_3+\ldots+\beta_{k-2}+\beta_{k-1}+\beta_k\right) &= R_1\\
\alpha_2\left(\beta_1+\beta_2+\beta_3+\ldots+\beta_{k-2}+\beta_{k-1}\right) &= R_2\\
\alpha_3\left(\beta_1+\beta_2+\beta_3+\ldots+\beta_{k-2}\right) &= R_3\\
&\ldots\\
\alpha_{k-2}\left(\beta_1+\beta_2+\beta_3\right) &= R_{k-2}\\
\alpha_{k-1}\left(\beta_1+\beta_2\right) &= R_{k-1}\\
\alpha_k\beta_1 &= R_k
\end{aligned} \tag{5}$$

And the second one for the columns of this table:

$$\begin{aligned}
\beta_1\left(\alpha_1+\alpha_2+\alpha_3+\ldots+\alpha_{k-2}+\alpha_{k-1}+\alpha_k\right) &= C_1\\
\beta_2\left(\alpha_1+\alpha_2+\alpha_3+\ldots+\alpha_{k-2}+\alpha_{k-1}\right) &= C_2\\
\beta_3\left(\alpha_1+\alpha_2+\alpha_3+\ldots+\alpha_{k-2}\right) &= C_3\\
&\ldots\\
\beta_{k-2}\left(\alpha_1+\alpha_2+\alpha_3\right) &= C_{k-2}\\
\beta_{k-1}\left(\alpha_1+\alpha_2\right) &= C_{k-1}\\
\beta_k\alpha_1 &= C_k
\end{aligned} \tag{6}$$

We can impose an additional restriction on the estimated parameters

$$\beta_1+\beta_2+\beta_3+\ldots+\beta_{k-2}+\beta_{k-1}+\beta_k=1. \tag{7}$$

This means we can interpret the β_j as a fraction of claims settled in the *j*-th development year. The equations (5), (6) and (7) will be solved iteratively in what follows.

According to (7) we can find α_1^1 the first approximation of α_1:

$$\alpha_1^1=R_1. \tag{8}$$

Now the first rough approximations of β_j can be counted according to (4):

$$\beta_j^1 = \frac{x_{1j}}{\alpha_1^1}, j = 1, 2, \ldots, k. \tag{9}$$

Next we find $\alpha_2^1, \ldots, \alpha_k^1$ from (5):

$$\alpha_i^1 = \frac{R_i}{\sum_{j=1}^{k-i+1} \beta_j^1}, i = 2, \ldots, k. \tag{10}$$

Further we compute the second iterations of all the estimated parameters - β_j^2 according to system (6):

$$\beta_j^2 = \frac{C_j}{\sum_{i=1}^{k-j+1} \alpha_i^1}, j = 1, 2, \ldots, k, \tag{11}$$

and α_i^2 from (5):

$$\alpha_i^2 = \frac{R_i}{\sum_{j=1}^{k-i+1} \beta_j^2}, i = 1, 2, \ldots, k. \tag{12}$$

In the next steps we repeat (11) and (12) with the up-to-date values – we get β_j^t from α_i^{t-1} and α_i^t from β_j^t for $i, j = 1, 2, \ldots, k$ and $t = 3, 4, \ldots, r$. The iterative procedure stops when the estimates in the *(r-1)*-th and *r*-th step do not differ significantly.

Let's compare the above described method to the well known chain ladder method (which is also derived for model (3)). Data from the exercise 3, [Kaas 2001], section 9.2, p. 218 will be used. A run-off triangle is given:

Table 3.

		Year of development				
		1	2	3	4	5
Year of origin	1	250,143	87,434	31,628	19,796	2,000
	2	293,227	102,494	37,075	23,205	**2,345**
	3	207,998	72,703	26,299	**16,460**	**1,663**
	4	318,628	111,372	**40,287**	**25,215**	**2,548**
	5	349,000	**121,988**	**44,127**	**27,619**	**2,790**

Source: own calculations.

Our task is to estimate the needed reserves. Using the model (3) together with the iterative method we get:

$$\hat{\alpha}_1 = 395,753 \quad \hat{\alpha}_2 = 463,916 \quad \hat{\alpha}_3 = 329,075 \quad \hat{\alpha}_4 = 504,103 \quad \hat{\alpha}_5 = 552,155$$
$$\hat{\beta}_1 = 0,632 \quad \hat{\beta}_2 = 0,221 \quad \hat{\beta}_3 = 0,080 \quad \hat{\beta}_4 = 0,050 \quad \hat{\beta}_5 = 0,005$$

When we complete the table with the estimates of the total paid amounts or total amounts to be paid in the future according to

$$\hat{X}_{i,j} = \hat{\alpha}_i \hat{\beta}_j ; \; i, j = 2, 3, 4, 5,$$

we get the numbers stated in the table 4.

Table 4.

		Development year				
		1	2	3	4	5
Year of origin	1	232	106	35	16	2
	2	258	115	56	27	
	3	221	82	4		
	4	359	71			
	5	349				

Source: own calculations.

The figures in boldface are the estimates of the amounts to be paid in the future, their sums are the required estimates of the reserves to be kept - res_c, for *c = 6, 7, 8 and 9*,

$$res_6 = 181,080 ;$$
$$res_7 = 71,005 ;$$
$$res_8 = 30,167 ;$$
$$res_9 = 2,790 .$$

The chain ladder method solution of the same example:

$$\hat{\alpha}_1 = 391,000 \quad \hat{\alpha}_2 = 458,347 \quad \hat{\alpha}_3 = 325,115 \quad \hat{\alpha}_4 = 498,043 \quad \hat{\alpha}_5 = 545,841$$
$$\hat{\beta}_1 = 0,64 \quad \hat{\beta}_2 = 0,224 \quad \hat{\beta}_3 = 0,081 \quad \hat{\beta}_4 = 0,051 \quad \hat{\beta}_5 = 0,005$$

Table 5.

		Development year				
		1	2	3	4	5
Year of origin	1	250,240	87,584	31,632	19,785	2,002
	2	293,342	102,670	37,080	23,192	**2,347**
	3	208,074	72,826	26,302	**16,451**	**1,665**
	4	318,748	111,562	**40,292**	**25,201**	**2,550**
	5	349,338	**122,268**	**44,159**	**27,620**	**2,795**

Source: own calculations.

This time the estimated reserves are

$$res_6 = 181,358;$$

$$res_7 = 71,025;$$

$$res_8 = 30,170;$$

$$res_9 = 2,795.$$

When we compare the results acquired by both the methods we can see that the estimates of the reserves are very similar. The chain ladder method is well known and ordinarily used. Our future plan is to explore the properties of the iterative method described above – its convergence, initial values sensitivity and possible other problems or advantages related to this method.

References

Anderson D. et al., *A practitioner's guide to generalized linear models – A foundation for theory, interpretation and application*, Watson Wyatt Worldwide, London 2005.

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.

UOGÓLNIONE MODELE LINIOWE W MATEMATYCE AKTUARIALNEJ

Streszczenie

Uogólnione modele liniowe (*Generalized Linear Models* − GLM) zapewniają ujednolicone teoretycznie i koncepcyjnie ramy dla wielu zagadnień w matematyce aktuarialnej, na przykład w analizie przeżycia, regresji logistycznej, modelach probitowych, złożonym rozkładzie Poissona, uogólnionej estymacji równań oraz w modelach wielopoziomowych. GLM rozszerzają modele liniowe w trojaki sposób:

- rozkład składnika losowego może pochodzić z rodziny rozkładów wykładniczych,
- wariancja wartości odpowiedzi jest określoną funkcją swojej średniej,
- funkcja łącząca nie musi być funkcją tożsamościową.

Takie sytuacje są typowe w praktyce ubezpieczeniowej. W statystycznych ramach GLM można otwarcie tworzyć założenia co do natury danych ubezpieczeniowych i ich związków ze zmiennymi prognostycznymi. Metody rozwiązywania GLM są bardziej efektywne pod względem technicznym, są bardziej eleganckie teoretycznie i wartościowe w praktyce. Modele te, dzięki statystykom diagnostycznym, pozwalają wybrać jedynie istotne zmienne, pozwalają również zweryfikować założenia modelu.

Podstawowym zastosowaniem modeli GLM w analizie ubezpieczeniowej jest ustalanie składki i *underwriting*. Warunki, które ograniczają możliwość dowolnego kształtowania składki (np. regulacje prawne), zwiększyły użyteczność GLM w celowej analizie marketingowej.

GLM są dobrze ugruntowane w teorii statystyki i oferują zakładom ubezpieczeniowym praktyczną metodę osiągania zadowalających zysków i przewagi konkurencyjnej. W tym artykule skupiono się na technikach określania IBNR jako ważnego zagadnienia w prognozowaniu całości roszczeń zaszłych, ale niezgłoszonych lub roszczeń zgłoszonych, ale w całości niespłaconych. Wiele tradycyjnych aktuarialnych metod określania rezerw okazuje się estymacją specjalnych przypadków GLM metodą maksymalizacji funkcji wiarygodności.