EKONOMETRIA ECONOMETRICS 4(46) • 2014

ISSN 1507-3866

Alicja Wolny-Dominik
Uniwersytet Ekonomiczny w Katowicach
e-mail: alicja.wolny-dominiak@ue.katowice.pl

JEDNOMODELOWA TARYFIKACJA *A PRIORI* W KRÓTKOTERMINOWYCH UBEZPIECZENIACH MAJĄTKOWYCH*

Streszczenie: W obecnej praktyce biznesowej zakłady ubezpieczeń majątkowych dla portfeli, w których występuje duża liczba polis (np. komunikacyjne, ubezpieczenie nieruchomości), wykorzystują w taryfikacji *a priori* dwa modele regresyjne: model częstości szkód (*claims freqiency*) oraz model wartości szkody (*claims severity*). Najczęściej stosowane są modele GLM oraz regresja negatywno-dwumianowa w przypadku nadmiernej dyspersji liczby szkód. Alternatywą dla dwumodelowej taryfikacji jest modelowanie wykorzystujące jedynie jeden model regresyjny, w którym zmienną objaśnianą jest łączna wartość szkód dla pojedynczej polisy przy założonym złożonym rozkładzie Poissona (*compound Poisson*). Taka taryfikacja wymaga zatem analizowania jedynie jednej grupy czynników ryzyka, pomija np. modelowanie liczby szkód. Ponadto pozwala na uwzględnianie nadmiernej liczby wartości zerowych, co jest bardziej zawiłe w przypadku taryfikacji dwumodelowej. Celem niniejszego artykułu jest zaproponowanie modelu regresyjnego w jednomodelowej taryfikacji uwzględniającego specyfikę danych w portfelu ubezpieczeniowym, tj. założenie niezależności polis w portfelu nie jest spełnione.

Słowa kluczowe: jednomodelowa taryfikacja, ubezpieczenia majątkowe, składka czysta, rozkład Tweedie, złożony rozkład Poissona.

DOI: 10.15611/ekt.2014.4.03

1. Wstęp

W obecnej praktyce biznesowej zakłady ubezpieczeń majątkowych dla portfeli, w których występuje duża liczba polis (np. komunikacyjne, ubezpieczenie nieruchomości), wykorzystują w taryfikacji *a priori* dwa modele regresyjne: model częstości szkód (*claim frequency model*) oraz model wartości pojedynczej szkody (*claim severity model*), wykorzystując informacje o liczbie szkód oraz wartości szkód dla poszczególnych polis [Dimakos, Di Rattalma 2002; Wolny-Dominiak 2011; Antonio, Valdez 2012]. Oba modele pozwalają na estymację wartości składki ryzyka (*pure*

* Praca częściowo finansowana przez grant Narodowego Centrum Nauki (nr NN 111461540).

risk premium) dla pojedynczego ryzyka rozumianej jako wartość oczekiwana łącznej wartości szkód dla pojedynczego ryzyka. W modelach tych zmiennymi objaśniającymi są wielokategorialne czynniki ryzyka ustalane osobno dla liczby szkód oraz wartości pojedynczej szkody. Najczęściej stosowane są modele GLM, tj. GLM--Poisson oraz GLM-gamma, ew. regresja negatywno-dwumianowa w przypadku nadmiernej dyspersji dla liczby szkód. Do estymacji parametrów modeli stosuje się metodę największej wiarygodności. Jako iż maksymalizacja funkcji wiarygodności nie może być przeprowadzona analitycznie (brak rozwiązania analitycznego dla równań skoringowych), niezbędne jest stosowanie algorytmów numerycznych. W pracy korzystamy z szybkiego algorytmu iteracyjnego IWSL [McCullagh, Wedderburn 1972].

Alternatywą dla dwumodelowej taryfikacji jest modelowanie wykorzystujące jedynie jeden model regresyjny, w którym zmienną objaśnianą jest łączna wartość szkód dla pojedynczej polisy przy założonym złożonym rozkładzie Poissona (compound Poisson) [Jørgensen, Paes De Souza 1994]. Jednomodelowa taryfikacja *a priori* jest alternatywą dla popularnej taryfikacji dwumodelowej głównie w sytuacji, gdy zakład ubezpieczeń dysponuje jedynie informacjami o łącznej wartości szkód dla indywidualnych polis, a nie zna liczby szkód N_i. Wymaga ona analizowania jedynie jednej grupy czynników ryzyka, ponadto uzyskujemy jeden błąd modelu, a nie dwa jak w przypadku taryfikacji dwumodelowej. Jednak problemem w podejściu jednomodelowym jest fakt, iż funkcji gęstości złożonego rozkładu Poissona nie można zapisać w postaci analitycznej. Wiadomo jednak powszechnie, iż jest on szczególnym przypadkiem rozkładu Tweedie dla $p \in (1, 2)$. Mimo iż rozkłady Tweedie również nie mają swojej analitycznej postaci, jednak należą do dyspersyjnej rodziny rozkładów, co pozwala na stosowanie algorytmu IWSL. Wadą tego podejścia jest jednak konieczność estymacji dodatkowego parametru p. Algorytm zaproponowany w pracy [Dunn, Smyth 2008] dla dużych zbiorów danych, a takimi są portfele masowe, działa bardzo wolno, co powoduje nieefektywność jego stosowania w praktyce.

Celem niniejszego artykułu jest zaproponowanie modelu regresyjnego w jednomodelowej taryfikacji, który uwzględnia specyfikę danych w portfelu ubezpieczeniowym w aspekcie niezależności ryzyk. W tym przypadku niezależność ta rozumiana jest jako niezależność zmiennych losowych reprezentujących łączną wartość szkód dla pojedynczego ryzyka. W pierwszej części artykułu przedstawiono podstawy teoretyczne związane z modelowaniem oraz estymacją w jednomodelowej taryfikacji. Druga część zawiera proponowany model klasy HGLM będący rozszerzeniem modelu GLM o efekty losowe. Artykuł kończy przykład empiryczny, który ma na celu zobrazowanie proponowanych modeli w procesie estymacji parametrów tych modeli. W obliczeniach wykorzystano program **R**.

2. Jednomodelowa taryfikacja *a priori*

Rozważmy portfel *n* polis w portfelach ubezpieczeń majątkowych o dużej liczbie polis. Każdej polisie[1] odpowiada pewna zmienna losowa o określonym rozkładzie, dalej oznaczana przez S_i, $i=1,...,n$. Oznaczmy przez N_i liczbę szkód dla i-tej polisy w portfelu, natomiast przez Y_{ik}, $k=1,...,N_i$ odpowiadającą jej wartość pojedynczej szkody[2]. Wtedy zmienna S_i ma postać:

$$S_i = Y_{i1} + ... + Y_{iN_i}$$

i określa łączną wartość szkód wygenerowaną przez *i*-tą polisę. Tak zdefiniowana zmienna, przy założeniach:

1) $N_i \sim Pois(\lambda_i)$,

2) $Y_{i1},...,Y_{iN_i}$ mają takie same rozkłady pochodzące z dyspersyjnej rodziny rozkładów wykładniczych z parametrami (μ_i, ϕ, p) spełniające warunek $Var(Y_{ik}) = \phi\mu_i^p$ (podrodzina Tweedie rozkładów),

3) $Y_{i1},...,Y_{iN_i}$ są niezależne oraz niezależne od N_i,

rozkłada się zgodnie ze złożonym rozkładem Poissona (ozn. *CPois*). W tym przypadku dwa pierwsze momenty rozkładu *CPois* mają następującą postać:

$$\begin{aligned} E[S_i] &= E[E[S_i \mid N_i]] = E[Y_i]E[N_i], \\ Var(S_i) &= E^2[Y_i]Var(N_i) + E[N_i]Var(Y_i). \end{aligned} \tag{1}$$

Składka ryzyka (*pure risk premium*) dla pojedynczego ryzyka definiowana jest jako:

$$\pi_i = E[S_i],\ i=1,...,n. \tag{2}$$

W celu uzyskania konkretnej wartości składki ryzyka w masowych portfelach polis zakłady ubezpieczeń powszechnie stosują odpowiednie modele statystyczne, w których wartość składki ryzyka jest estymowaną wartością $\hat{\pi}_i$ na podstawie informacji zawartych w portfelu (próbie statystycznej). Jako iż charakterystyczną cechą portfela polis jest jego niejednorodność, powodująca generowanie różnych wartości szkód dla polisy, stosowane są najczęściej modele regresyjne klasy GLM. Portfel różnicują czynniki ryzyka charakteryzujące ogólnie osobę ubezpieczającą się, przedmiot ubezpieczenia oraz zmienną przestrzenną (w sensie geograficznym).

[1] W kontekście ubezpieczeniowym polisa wraz z odpowiadającymi jej zmiennymi losowymi nazywana jest ryzykiem, natomiast portfel polis – portfelem ryzyk.

Przyjmijmy w rozważanym portfelu założenie niezależności zmiennych $S_1,...,S_n$. Wtedy model ma postać:

$$\begin{cases} S_i \sim CPois(\mu_i, \phi, p) \\ \mu_i = E_i \exp(\boldsymbol{\beta}^T \mathbf{x}_i) \end{cases}, \tag{3}$$

gdzie $\boldsymbol{\beta} = (\beta_0, \beta_1,..., \beta_k)^T$ jest wektorem efektów stałych, $\mathbf{x}_i$ jest *i*-tym wierszem macierzy modelu **X,** natomiast E_i oznacza ekspozycję na ryzyko i jest to najczęściej czas trwania polisy. Wektor parametrów modelu ma zatem postać $(\beta_0, \beta_1,..., \beta_k, p, \phi)^T$. Korzystając z tego, iż $\boldsymbol{\beta}^T$ jest wektorem stałych, wartość składki ryzyka dla i-tej polisy wynosi:

$$\hat{\pi}_i = E_i \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i). \tag{4}$$

W estymacji parametrów modelu (3) zastosowanie znajduje metoda największej wiarygodności. W pracy [Jørgensen, Paes De Souza 1994] autorzy wykazali, iż złożony rozkład Poissona z przyjętym rozkładem gamma elementów sumy jest specyficznym przypadkiem rozkładu Tweedie, w którym $1 < p < 2$. Dzięki temu, mimo że ten przypadek złożonego rozkładu Poissona nie ma swojej analitycznej postaci funkcji gęstości, znana jest jego postać funkcji wariancji $V(\mu_i)$. Pozwala to zatem na numeryczne wyznaczenie estymatorów parametrów modelu z wykorzystaniem klasycznego algorytmu IWSL bez potrzeby znajomości postaci funkcji wiarygodności.

3. Model z efektami stałymi i losowymi klasy HGLM

Zauważmy, iż założenie niezależności zmiennych $S_1,...,S_n$ nie zawsze jest spełnione w portfelu ryzyka, a spełniona jest jedynie niezależność warunkowa. Analizując ubezpieczenie domu, należy pamiętać, że w momencie, kiedy w jednym domu wybuchnie pożar, to pożar może wybuchnąć również w domu stojącym obok, natomiast nie wybuchnie w domu stojącym w dużej odległości [Otto 2013]. Fakt ten można uwzględnić, przechodząc od modelu GLM do modelu GLM z efektami losowymi $(u_1,...,u_K)^T$, uzyskując model mieszany, gdzie u_i, $i = 1,...,K$ są niezależnymi realizacjami zmiennej losowej U. Oznacza to podział portfela na klastry spełniające warunek: polisy należące do tego samego klastra są zależne, natomiast dwie polisy z dwóch różnych klastrów są niezależne. Zatem spełniony jest warunek:

$$\operatorname{cov}(S_{ij}, S_{ik} | U) \neq 0, \; j,k = 1,...,n_i, \tag{5}$$

gdzie S_{ij}, S_{ik} oznaczają łączną wartość szkód dla *j*-tej oraz *k*-tej polisy należących do *i*-tego klastra, natomiast n_i oznacza liczebność *i*-tego klastra. W przypadku gdy

założymy, iż niezależne efekty losowe u_i przyjmują rozkład z dyspersyjnej rodziny wykładniczej (ozn. EDM – *Exponential Dispersion Model* [Jorgensen 1987]) z parametrami μ_u, ϕ_u, uzyskujemy model klasy HGLM [Lee, Nelder 1996]. Postać modelu jest następująca:

$$\begin{cases} S_{ij} \mid U \sim CPois(\mu_{ij}, \phi, p) \\ u_i \sim EDM(\mu_u, \phi_u) \\ \mu_{ij}(u) = E_{ij} \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij} + v(u)^T z_{ij}) \end{cases}, \tag{6}$$

gdzie $v(u) = (\ln(u_1),...,\ln(u_K))^T$ [3]. Wektor parametrów modelu wynosi wtedy $(\beta_0, \beta_1,..., \beta_r, \phi, \mu_u, \phi_u)^T$. Wartość składki ryzyka dla *i*-tej polisy może w tym przypadku być przyjmowana dwojako:

$$\begin{aligned} \hat{\pi}_{ij}(u) &= E[S_{ij} \mid U] = E_{ij} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij} + v(u)^T z_{ij}) \\ \hat{\pi}_{ij} &= E[S_{ij}] = E[E[S_{ij} \mid U]] = E_{ij} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij}) E[v(\hat{u})^T z_{ij}] = \\ &= E_{ij} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ij}) \hat{\mu}_u. \end{aligned} \tag{7}$$

Estymacja parametrów modelu (6) nie jest już taka oczywista jak w przypadku modelu (3). Wynika to z faktu, iż znane są postaci rozkładów dla rozkładu warunkowego zmiennej $S_{ij} \mid U$ (Tweedie z parametrem $1 < p < 2$) oraz efektów losowych $u_1,...,u_K$, natomiast nie jest znana postać rozkładu brzegowego zmiennej S_{ij}. W takiej sytuacji można wykorzystać funkcję rozszerzonej wiarygodności będącą *de facto* rozszerzeniem funkcji *quasi*-wiarygodności [Bjørnstad 1996]. W pracy korzystamy z pewnego przypadku funkcji rozszerzonej wiarygodności zwanej funkcją hierarchicznej wiarygodności i oznaczanej przez $H(.)$ [4]:

$$H(\beta_0, \beta_1,..., \beta_r, \phi, \mu_u, \phi_u; s_{ij}) = \prod_{i=1}^{K} \prod_{j=1}^{n_i} f_{\mu,\phi}(s_{ij} \mid u) g_{\mu_u,\phi_u}(u). \tag{8}$$

Do wyznaczenia wartości maksymalnych funkcji log-H(.) zastosowanie znajduje algorytm iteracyjny H-IWSL (hierarchiczny IWSL) [Lee, Nelder 1996]. Algorytm ten zawiera swoją implementację w programie R w pakiecie {hglm} [Ronnegard i in. 2010].

[3] Opis przekształcenia *v*(.) por. [Lee i in. 2006 s. 102], Example 4.3.

[4] Warunkiem tego, aby funkcja rozszerzonej wiarygodności była funkcją hierarchicznej wiarygodności, zob. [Lee i in. 2006, s. 175-176].

4. Przykład obliczeniowy

W celu zobrazowania działania modelu HGLM w jednomodelowej taryfikacji analizujemy portfel 664 polis. Został on zaczerpnięty z pracy [Ohlsson i Johansson 2010] i zawarty w pakiecie programu R o nazwie {`insuranceData`} [Wolny-Dominiak, Trzęsiok 2014]. Czynniki ryzyka zarejestrowane w portfelu przedstawia tab. 1.

Tabela 1. Opis czynników ryzyka w portfelu

Nazwa czynnika ryzyka	Opis zmiennej	Kategorie zmiennej
Wiek.k	Wiek kierowcy	Kategorie A-G (najstarszy)
Klasa.MC	Współczynnik: moc silnika dzielona przez wagę pojazdu + 75(kg)	Kategorie A-G
Wiek.poj	Wiek pojazdu	Kategorie A-C
Region	Region użytkowania pojazdu	Kategorie A-G

Źródło: obliczenia własne.

Przyjmujemy, iż portfel pogrupowany jest na klastry ze względu na czynnik *Region*, który najczęściej generuje zależności w portfelu (o tzw. *area effect* por. [Dimakos, Di Rattalma 2002; Żądło 2014]). Zatem w modelu mamy $u_1,...,u_7$ efektów losowych. Przyjmując klasycznie rozkład normalny dla zmiennej geograficznej (wtedy $\phi_u = \sigma_u^2$), należy zauważyć, że postać modelu jest następująca:

$$\begin{cases} S_{ij} \mid U \sim CPois(\mu_{ij}, \varphi, p) \\ u_i \sim N(0, \sigma_u^2) \\ \mu_{ij}(u) = E_{ij} \exp[\beta_0 + \boldsymbol{\beta}_1 Wiek.k_{ij} + \boldsymbol{\beta}_2 Klasa.MC_{ij} + \boldsymbol{\beta}_3 Wiek.poj_{ij} + \\ + v(REgion_{ij})] \end{cases}, \tag{9}$$

gdzie $i = 1,...,7$, $j = 1,...,n_i$, $n_1 + ... + n_7 = 664$. Ze względu na to, że każdy czynnik w modelu jest zmienną wielokategorialną, każdy wektor parametrów $\boldsymbol{\beta}_i$ posiada tyle współrzędnych, ile jest kategorii danego czynnika, np. $\boldsymbol{\beta}_1 = (\beta_{1A}, \beta_{1B},..., \beta_{1G})^T$. Tabela 2 przedstawia uzyskane szacunki parametrów modelu (9), które można dalej wykorzystać do wyznaczenia wartości składki dla pojedynczego ryzyka. Dodatkowo tab. 2 zawiera oszacowania parametrów w modelu GLM, bez efektów losowych u_i.

Tabela 2. Szacunkowe wartości parametrów modeli GLM oraz HGLM

		GLM		HGLM	
		$\exp(\hat{\beta})$	s.e.	$\exp(\hat{\beta})$	s.e.
Wyraz wolny	β_0	24 625	0,4708	27 493	0,3562
Wiek.kierB	β_{1B}	0,9633	0,3853	0,9685	0,2801
Wiek.kierC	β_{1c}	0,9431	0,3854	0,9457	0,2813
Wiek.kierD	β_{1D}	0,9303	0,4131	0,8931	0,3004
Wiek.kierE	β_{1E}	0,4006	0,3824	0,4040	0,2769
Wiek.kierF	β_{1F}	0,5016	0,4084	0,5549	0,2962
Wiek.kierG	β_{1G}	0,4003	0,6071	0,3903	0,4394
Klasa.MCB	β_{2B}	0,4521	0,4166	0,4100	0,3012
Klasa.MCC	β_{2c}	0,5809	0,3535	0,5111	0,2560
Klasa.MCD	β_{2D}	0,9561	0,3822	0,7949	0,2765
Klasa.MCE	β_{2E}	0,5378	0,3571	0,4766	0,2585
Klasa.MCF	β_{2F}	0,8178	0,3518	0,7408	0,2551
Klasa.MCG	β_{2G}	**1,5738**	**0,9754**	**1,7220**	**0,7071**
Wiek.pojB	β_{3B}	0,5864	0,2601	0,5626	0,1880
Wiek.pojC	β_{3C}	0,2436	0,2206	0,2395	0,1596

Źródło: obliczenia własne.

Rozkład łącznej wartości szkód ma dodatkowe parametry ϕ, p, których szacunki w naszym przypadku wynoszą $\hat{\phi} = 2{,}47$ oraz $\hat{p} = 1{,}99$. Z kolei parametr rozkładu efektów losowych $\hat{\sigma}_u^2 = 0{,}0429$. Zauważmy, iż generalnie model HGLM generuje mniejsze błędy dla parametrów w stosunku do modelu GLM. Największy błąd występuje dla parametru β_{2G}, co jest spowodowane tym, iż w portfelu występuje tylko jedna polisa o takiej kategorii zmiennej *Klasa.MCG*.

5. Zakończenie

Wprowadzenie do modelu GLM efektów losowych uwzględniających zależności w portfelu nie jest nowością w omawianej tematyce. W pracy [Nelder, Verrall 1997] autorzy przedstawiali poszczególne modele wiarygodności jako szczególne przypadki modelu HGLM, gdzie efektami losowymi są czynniki nieobserwowalne charakte-

rystyczne dla indywidualnych ryzyk (np. cechy charakteru osoby ubezpieczającej się). Z kolei w pracy [Dimakos, Di Rattalma 2002] autorzy proponowali model z losowymi efektami przestrzennymi, którego parametry szacowali w ujęciu bayesowskim. W proponowanym w pracy modelu HGLM zależności występujące w portfelu ryzyk dotyczą pewnych zjawisk, które charakteryzują dana grupę ryzyk (klaster). To znaczy, że polisy w klastrze dzielą jeden wspólny efekt losowy, w odróżnieniu np. od efektów indywidualnych. Jest to zatem podejście zbliżone do tego zaproponowanego w pracy [Dimakos, Di Rattalma 2002]. Jednak w odróżnieniu od podejścia bayesowskiego stosujemy podejście klasyczne, bazujące ma metodzie największej wiarygodności. Wydaje się ono bardziej intuicyjne dla praktyków na co dzień stosujących modele GLM.

Literatura

Antonio K., Valdez E.A., 2012, *Statistical concepts of a priori and a posteriori risk classification in insurance*, AStA Advances in Statistical Analysis, 96(2), 187-224.

Bjørnstad J.F., 1996, *On the generalization of the likelihood function and the likelihood principle*, Journal of the American Statistical Association, 91(434), 791-806.

De Jong, P., Heller G.Z., 2008, *Generalized linear models for insurance data* (Vol. 136), Cambridge University Press, Cambridge.

Dimakos X.K., Di Rattalma A.F., 2002, *Bayesian premium rating with latent structure*, Scandinavian Actuarial Journal, 2002(3), 162-184.

Dunn P.K., Smyth G.K., 2008, *Evaluation of Tweedie exponential dispersion model densities by Fourier inversion*, Statistics and Computing, 18(1), 73-86.

Jorgensen B., 1987, *Exponential dispersion models*, Journal of the Royal Statistical Society. Series B (Methodological), 127-162.

Jørgensen B., Paes De Souza M.C., 1994, *Fitting Tweedie's compound Poisson model to insurance claims data*, Scandinavian Actuarial Journal, 1994(1), 69-93.

Lee Y., Nelder J.A., 1996, *Hierarchical generalized linear models*, Journal of the Royal Statistical Society, Series B (Methodological), 619-678.

Lee Y., Nelder J.A., Pawitan Y., 2006, *Generalized linear models with random effects: unified analysis via H-likelihood*, CRC Press.

McCullagh P., Wedderburn R.W.M., 1972, *Generalized linear model*, Journal of the Royal Statistical Society. Series A (General), Vol. 135/3, 370-384.

Nelder J.A., Verrall R.J., 1997, *Credibility theory and generalized linear models*, Astin Bulletin 27.01: 71-82.

Ohlsson E., Johansson B., 2010, *Non-life Insurance Pricing with Generalized Linear Models*, Springer.

Otto W., 2013, *Ubezpieczenia majątkowe. Część I. Teoria ryzyka*, Wydawnictwo WNT.

Ronnegard L., Xia Shen, Moudud A., 2010, *hglm: a package for fitting hierarchical generalized linear models*, The R Journal, 2(2), 20-28.

Wolny-Dominiak A., Trzęsiok M., 2014, *insuranceData: A Collection of Insurance Datasets Useful in Risk Classification in Non–life Insurance*, R package version 1.0 http://CRAN.R–project.org/package=insuranceData.

Wolny-Dominiak A., 2011, *Szacowanie stóp taryf w ubezpieczeniach majątkowych z wykorzystaniem modelu HGLM*, Zeszyty Naukowe/Uniwersytet Ekonomiczny w Poznaniu, (182), 318-328.

Żądło T., 2014, *On longitudinal moving average model for prediction of subpopulation total*, Statistical Papers, 1-23.

SINGLE-MODEL A PRIORI RATEMAKING IN SHORT TERM NON-LIFE INSURANCE

Summary: The goal of this paper is to propose the regression model usefull in a priori ratemaking in short term non-life insurance. In the model the aggregat claim amount for individual risk following is estimated. It is asumed that this random variable following the compound Poisson distribution being a special case of Tweedie. We notice that the independent assumtion in the portfolio of risks is violated. That is why we adopt the mixed model with fixed and random effects in place of the model with fixed effects only. In the first part of the paper the theoretical model is presented while in the second part practical application is analised. All calculations in the case study are made in R software.

Keywords: single-model ratemaking, non-life insurance, pure risk premium, Tweedie, compound Poisson.