Zagadnienia aktuarialne – teoria i praktyka

pod redakcją
Walentego Ostasiewicza

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2011

Recenzenci
Helena Jasiulewicz, Wojciech Kordecki, Zbigniew Palmowski

Redaktor Wydawnictwa
Aleksandra Śliwka

Redakcja techniczna
Barbara Łopusiewicz

Korektor
Barbara Cibis

Łamanie
Beata Mazur

Projekt okładki
Beata Dębska

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192
ISBN 978-83-7695-186-7

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 207
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS
Zagadnienia aktuarialne – teoria i praktyka ISSN 1899-3192

Alicja Wolny-Dominiak
Uniwersytet Ekonomiczny w Katowicach

ANALIZA PORÓWNAWCZA MODELI MIESZANYCH SZACOWANIA STÓP TARYF W UBEZPIECZENIACH MAJĄTKOWYCH Z WYKORZYSTANIEM KROSWALIDACJI

Streszczenie: W pracy przedstawiono zastosowanie hierarchicznych mieszanych modeli liniowych (HGLM) w procesie wyznaczania stóp taryf *a priori*. Jako zmienną objaśnianą przyjęto średnią wartość wypłacanych odszkodowań (zmienna ciągła), a jako zmienną objaśniającą – zmienne taryfikacyjne charakteryzujące przedmiot ubezpieczenia oraz osobę ubezpieczającą się (zmienne nominalne wielokategorialne). W modelach HGLM istotnym problemem obliczeniowym jest wyznaczenie dewiancji, która może być stosowana jako kryterium wyboru ostatecznej postaci modelu. Zatem w celu uproszczenia początkowej fazy obliczeniowej w taryfikacji *a priori* (pomijając wyznaczanie dewiancji), do oceny modelu zaproponowano wykorzystanie procedury *k*-krotnej kroswalidacji. Analizę przypadku przedstawiono dla portfela polis komunikacyjnych. Procedura zaimplementowana została w programie komputerowym R.

Słowa kluczowe: stopa taryfy, hierarchiczny mieszany model liniowy, kroswalidacja.

1. Wstęp

Zakłady ubezpieczeń majątkowych powszechnie stosują w procesie wyznaczania stóp taryf *a priori* uogólnione modele liniowe (GLM), w których średnią (całkowitą) wartość szkody (zmienna objaśniana) uzależnia się od ustalonych zmiennych taryfikacyjnych (zmienne objaśniające). W modelach GLM przyjmuje się, iż wszystkie zmienne objaśniające mają stały wpływ na zmienną objaśnianą (*fixed effect*). Rozszerzeniem modeli GLM są liniowe modele mieszane, w których zakłada się, że część zmiennych taryfikacyjnych ma losowy wpływ na zmienną objaśnianą (*random effects*), co oznacza, że parametry strukturalne dla tych zmiennych są zmiennymi losowymi o określonych rozkładach […]. W przypadku modelowania stóp taryf w ubezpieczeniach majątkowych zastosowanie znajdują mieszane modele liniowe klasy HGLM (*Hierarchical Generalized Linear Models*). W modelach HGLM losowe parametry strukturalne mogą przyjmować rozkład z dyspersyjnej rodziny rozkładów wykładniczych. To elastyczne założenie ma zasadnicze znaczenie w modelowaniu danych ubezpieczeniowych, gdzie rozkład normalny znajduje zastosowanie rzadko, częściej rozkład gamma lub rodzina rozkładów Tweedie, czyli takich, dla których zachodzi $V(\mu) = \mu^p$ [Lee i in. 2006]. Mimo zalet stosowania modeli HGLM, które daje możliwość modelowania dyspersji, ich istotną wadą praktyczną jest stopień złożoności w procesie szacowania parametrów strukturalnych, mierników dobroci dopasowania, jak również walidacja modeli. W pracy przedstawiono procedurę *k*-krotnej kroswalidacji pozwalającą na szybką preselekcję modelu HGLM bez konieczności estymacji np. dewiancji.

W pierwszej części artykułu przedstawiono teoretyczny ogólny model HGLM do szacowania stóp taryf *a priori* typu gamma-gamma oraz proponowaną procedurę kroswalidacji. W części zasadniczej artykułu przedstawiono empiryczną analizę modeli na danych rzeczywistych zaczerpniętych z literatury przedmiotu [de Jong, Heller 2008]. Zastosowano model HGLM w trzech scenariuszach: uwzględniając wpływ zmiennych taryfikacyjnych na zmienną objaśnianą, uwzględniając dodatkowo interakcje pomiędzy zmiennymi, uwzględniając wpływ zmiennej taryfikacyjnej na dyspersję modelu. Wykorzystano funkcję `hglm{hglm}` oraz przeprowadzono 10-krotną kroswalidację, korzystając z programu komputerowego R.

2. Model HGLM do szacowania stóp taryf w ubezpieczeniach komunikacyjnych

Proces wyznaczania stóp taryf w ubezpieczeniach komunikacyjnych można zapisać w postaci modelu HGLM. Niech Y będzie średnią wartością szkody w portfelu, X_i zmienną taryfikacyjną mającą stały wpływ na wartości szkód, $i = 1, \ldots, I$ zaś Z zmienną taryfikacyjną o losowym wpływie na wartości szkód. Zmienna objaśniana jest zmienną ciągłą, natomiast zmienne objaśniające są ciągłe lub nominalne wielokategorialne. Model HGLM typu gamma-gamma ma ogólną postać [Lee i in. 2006]:

$$\eta = g(\mu) = \mathrm{X}\beta + \mathrm{Z}u,$$

gdzie $\beta = [\beta_1, ..., \beta_K]$ oraz $u = [u_1, ..., u_K]$, **X** oraz **Z** są macierzami modelu. Parametry strukturalne modelu mają następującą interpretację: parametr β_i, $i = 1, \ldots, I$, mierzy wpływ *i*-tej zmiennej taryfikacyjnej na wartość szkód (stały dla wszystkich kategorii), a parametr u_k, $k = 1, ..., K$, mierzy poziom ryzyka w obrębie kategorii (zmienny dla każdej kategorii). Zmienna $Y|u$ oraz zmienna u mają rozkłady gamma.

Oszacowane wartości parametrów strukturalnych służą do wyznaczenia wskaźników taryfikacyjnych t. Wskaźniki taryfikacyjne w modelu gamma-gamma, w którym występują logarytmiczne funkcje połączenia, można przedstawić następująco:

$$t_i = \exp(\beta_i),\ i = 1, ..., I,\ t_k = \exp(u_k),\ k = 1, ..., K.$$

Iloczyn wskaźników taryfikacyjnych dla wszystkich zmiennych uwzględnianych w modelu daje ostateczną stopę taryfy, dzieląc portfel niejednorodny na subportfele jednorodne.

Następnie dla konkretnej polisy wyznaczyć można stopę taryfy uzależnioną od liczby zmiennych taryfikacyjnych wprowadzonych do modelu. Składkę bazową można przyjąć np. na poziomie wyznaczonym przez wyraz wolny dla modelu GLM zmiennej $Y|u$.

3. Procedura *k*-krotnej kroswalidacji dla modelu HGLM

W pracy [Lee i in. 2006] zaproponowano zmodyfikowaną wersję dewiancji dla modeli HGLM pozwalającą na wybór z grupy modeli modelu najlepszego, czyli takiego, który daje minimalną wartość dewiancji, podobnie jak np. kryterium AIC. Niestety, miara ta w przypadku szacowania parametrów funkcją hierarchicznej wiarygodności (a nie klasycznej jak w przypadku modeli GLM) jest skomplikowana w szacowaniu.

Alternatywą dla dewiancji może być szacowanie rzeczywistego błędu predykcji dla modelu HGLM z wykorzystaniem *k*-krotej kroswalidacji [Burman 1989]. W niniejszej pracy zastosowano następujący algorytm dla $k = 10$:

a) losowe wyznaczenie ze zbioru danych 10 podzbiorów o zbliżonej liczebności (n – liczebność całego zbioru, m_l – liczebność l-tego podzbioru, $l = 1,...,10$),

b) 10-krotne szacowanie modelu HGLM na podzbiorze danych o liczebności $n - m_l$ z usunięciem zbioru walidującego,

c) 10-krotne wyznaczenie błędu $MSE_l = \dfrac{\sum (y - \hat{\mu}_l)^2}{m_l}$,

d) szacowanie błędu kroswalidacji: $cv = \sum_{l=1}^{10} \dfrac{m_l}{n} MSE_l$.

Porównując modele, wybrano model o najmniejszej wartości *cv*. Implementację procedury w programie komputerowym R przedstawiono w załączniku A.

4. Przykład empiryczny

Model HGLM zastosowano do budowy taryf *a priori* w ubezpieczeniach komunikacyjnych. Wykorzystano bazę danych szkód komunikacyjnych (*third party motor insurance claims*) zaczerpniętą z bazy `http://www.acst.mq.edu.au/research/books/GLMsforInsuranceData/data_sets` [de Jong, Heller 2008]. Baza danych zawiera następujące dane uwzględnione w modelu:

a) Zmienne taryfikacyjne:

– *veh_value* – wartość pojazdu,
– *veh_age* – wiek pojazdu: 1 (najmłodszy), 2, 3, 4,
– *gender* – płeć kierowcy: 0 (kobieta), 1,

– *agecat* – wiek kierowcy: 1 (najmłodszy), 2, 3, 4, 5, 6,
– *veh_body* – kształt pojazdu.

b) Zmienna objaśniana:

– *claimcst0* – wartość szkody.

c) Ekspozycja na ryzyko (mierzona w okresie trwania polisy w stosunku do całego okresu uwzględnianego w bazie):

– *exposure* – wartości z przedziału [0,1] .

W pierwszym kroku rozpatrywano trzy modele HGLM typu gamma-gamma, przyjmując jako efekt losowy zmienną *veh_body*:

1. Model 1 – średnia wartość szkody uzależniona jest od zmiennych taryfikacyjnych, takich jak: wartość pojazdu, wiek pojazdu, płeć kierowcy, wiek kierowcy, kształt pojazdu

$$claimcst0 = \beta_0^1 + \beta_1^1 veh\_value + \beta_2^1 veh\_age +$$
$$+\beta_3^1 gender + \beta_4^1 agecat + u_1^1 veh\_body.$$

2. Model 2 – średnia wartość szkody uzależniona jest od zmiennych taryfikacyjnych, takich jak: wartość pojazdu, wiek pojazdu, płeć kierowcy, wiek kierowcy, kształt pojazdu, oraz występuje interakcja pomiędzy płcią i wiekiem kierowcy

$$claimcst0 = \beta_0^2 + \beta_1^2 veh\_value + \beta_2^2 veh\_age + \beta_3^2 gender +$$
$$+\beta_4^2 agecat + \beta_5^2 gender * agecat + u_1^2 veh\_body.$$

3. Model 3 – średnia wartość szkody uzależniona jest od zmiennych taryfikacyjnych, takich jak: wartość pojazdu, wiek pojazdu, płeć kierowcy, wiek kierowcy, kształt pojazdu, a parametr dyspersji uzależniony jest od wieku pojazdu

$$claimcst0 = \beta_0^3 + \beta_1^3 veh\_value + \beta_2^3 gender + \beta_3^3 agecat + u_1^3 veh\_body$$
$$\phi = \gamma_0 + \gamma_1 veh\_age.$$

W modelach przyjęto założenie, iż obserwowane wartości szkód y_i są niezależne, oraz przyjęto tożsamościową funkcję połączenia w modelu dyspersji. Oszacowane parametry strukturalne modelu przedstawiają tabele 1-3.

Tabela 1. Stałe parametry strukturalne w modelu 1

	Parametry stałe	Średni błąd szacunku	*p*-wartość
$\hat{\beta}_0^1$	7,5908	0,1561	0,0000
veh_value	0,0282	0,0249	0,2586
veh_age	0,0631	0,0239	0,0082
Gender	0,1786	0,0387	0,0000
Agecat	–0,0631	0,0132	0,0000

Źródło: obliczenia własne.

Tabela 2. Stałe parametry strukturalne w modelu 2

	Parametry stałe	Średni błąd szacunku	p-wartość
$\hat{\beta}_0^2$	7,5480	0,1596	0,0000
veh_value	0,0261	0,0249	0,2964
veh_age	0,0619	0,0239	0,0096
Gender	0,2837	0,0963	0,0032
Agecat	–0,0481	0,0179	0,0073
gend:agecat	–0,0312	0,0262	0,2340

Źródło: obliczenia własne.

Tabela 3. Stałe parametry strukturalne w modelu 3

	Parametry stałe	Średni błąd szacunku	p-wartość
$\hat{\beta}_0^3$	7,7411	0,1265	0,0000
veh_value	–0,0079	0,0186	0,6722
Gender	0,1840	0,0387	0,0000
Agecat	–0,0624	0,0132	0,0000

Źródło: obliczenia własne.

We wszystkich modelach p-wartość pokazuje, iż zmienna *veh_value* jest statystycznie nieistotna na poziomie istotności 5% oraz nie występuje interakcja pomiędzy zmiennymi *gender* oraz *agecat*. Do dalszej analizy zmodyfikowano więc modele następująco:

4. Model 1.1 – średnia wartość szkody uzależniona jest od zmiennych taryfikacyjnych, takich jak: wiek pojazdu, płeć kierowcy, wiek kierowcy, kształt pojazdu:

$$claimcst0 = \beta_0^{1.1} + \beta_1^{1.1} veh\_age + \beta_2^{1.1} gender + \beta_3^{1.1} agecat + u_1^{1.1} veh\_body.$$

5. Model 3.1 – średnia wartość szkody uzależniona jest od zmiennych taryfikacyjnych, takich jak: wiek pojazdu, płeć kierowcy, wiek kierowcy, kształt pojazdu, a parametr dyspersji uzależniony jest od wieku pojazdu

$$claimcst0 = \beta_0^{3.1} + \beta_1^{3.1} gender + \beta_2^{3.1} agecat + u_1^{3.1} veh\_body$$
$$\phi = \gamma_0 + \gamma_1 veh\_age.$$

Oszacowane parametry strukturalne zmodyfikowanych modeli przedstawiają tab. 4-5.

Tabela 4. Stałe parametry strukturalne w modelu 1.1

	Parametry stałe	Średni błąd szacunku	*p*-wartość
$\hat{\beta}_0^{1.1}$	7,6998	0,1237	0,0000
Veh_age	0,0455	0,0178	0,0105
gender	0,1800	0,0387	0,0000
agecat	–0,0640	0,0132	0,0000

Źródło: obliczenia własne.

Tabela 5. Stałe parametry strukturalne w modelu 3.1

	Parametry stałe	Średni błąd szacunku	*p*-wartość
$\hat{\beta}_0^{3.1}$	7,7202	0,1210	0,0000
gender	0,1829	0,0387	0,0000
agecat	–0,0613	0,0132	0,0000

Źródło: obliczenia własne.

Widać, iż w powyższych modelach wszystkie zmienne są statystycznie istotnie.

W drugim kroku przeprowadzono 10-krotną kroswalidację, uzyskując następujące miary *cv* dla obu modeli:

– $cv^{1.1} = 12667.022$,
– $cv^{3.1} = 12649.699$.

Zatem model 3.1 jest lepszy. W modelu tym losowe parametry dla kształtu samochodu przedstawiają się następująco:

Tabela 6. Losowe parametry strukturalne modelu 3.1

Veh_body	Parametry losowe	Średni błąd szacunku	Wskaźniki taryfikacyjne
BUS	–0,146	0,268	0,86
CONVT	0,044	0,310	1,05
COUPE	0,238	0,167	1,27
HBACK	0,043	0,117	1,04
HDTOP	0,055	0,146	1,06
MCARA	–0,514	0,246	0,60
MIBUS	0,222	0,186	1,25
PANVN	0,037	0,171	1,04
RDSTR	–0,117	0,321	0,89
SEDAN	–0,064	0,116	0,94
STNWG	–0,032	0,116	0,97
TRUCK	0,182	0,148	1,20
UTE	0,050	0,131	1,05

Źródło: obliczenia własne.

Składka bazowa wyznaczana jako eksponent wyrazu wolnego modelu 3.1 wynosi 2253,37. Zatem skorygowana składka bazowa dla kształtu pojazdu, np. COUPE, wynosi 2858,19. Następnie w celu wyznaczenia ostatecznej stopy taryfy należy dalej korygować składkę bazową o kolejne wskaźniki taryfikacyjne dla innych zmiennych taryfikacyjnych występujących w modelu 3.1.

5. Podsumowanie

W artykule przedstawiono w zarysie ideę hierarchicznych liniowych modeli mieszanych oraz możliwości aplikacyjne w wyznaczaniu wskaźników taryfikacyjnych oraz budowie taryf. Przeanalizowano trzy modele, modelując dane zawarte w rzeczywistej bazie danych komunikacyjnych. Wyboru postaci modelu dokonano, stosując procedurę 10-krotnej kroswalidacji. Przeprowadzona procedura pokazuje, w jaki sposób można łatwo dokonać wstępnego wyboru modelu HGLM. W celu ostatecznego przyjęcia modelu do wyznaczania stóp taryf *a priori* niezbędne jest przeprowadzenie weryfikacji statystycznej modelu.

Literatura

Burman P. (1989), *A comparative study of ordinary cross-validation,* v-*fold cross-validation and repeated learning-testing methods*, „Biometrika" no 76.

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

Lee Y., Nelder A.J., Pawitan Y. (2006), *Generalized Linear Models with Random Effects, Monographs on Statistics and Applied Probability 106*, Chapman&Hall\CRC.

Załącznik A. Implementacja procedury kroswalidacji w programie komputerowym R

```
motor=read.csv(file="car.csv")

h1=hglm(fixed=claimcst0~veh_value+veh_
age+gender+agecat,random=~1|veh_body, family=Gamma(lwejsciek="log"),
rand.family=Gamma(lwejsciek="log"),data=motor)

h2=hglm(fixed=claimcst0~veh_value+veh_age+gender+agecat+gender*agecat,
random=~1|veh_body, family=Gamma(lwejsciek="log"), rand.family=Gamma
(lwejsciek="log"),data=motor)

h3=hglm(fixed=claimcst0~veh_value+gender+agecat,random=~1|veh_body,
family=Gamma(lwejsciek="log"),disp=~veh_age, data=motor)

h1.1=hglm(fixed=claimcst0~veh_age+gender+agecat,random=~1|veh_body,
family=Gamma(lwejsciek="log"),
rand.family=Gamma(lwejsciek="log"),data=motor)
```

```
h3.1=hglm(fixed=claimcst0~gender+agecat,random=~1|veh_body,
family=Gamma(lwejsciek="log"),disp=~veh_age, data=motor)

K=10
n <- nrow(motor)
wyjscie <- NULL
K <- round(K)
K1 <- unique(round(n/(1L:floor(n/2))))
temp <- abs(k1 - K)
f <- ceiling(n/K)
s <- sample(rep(1L:K, f), n)
n.s <- table(s)
y <- motor$claimcst0
mse.01 <- mean((y-h1$fv)^2)
mse.03 <- mean((y-h3$fv)^2)
ms <- max(s)
CV1 <- 0
CV3 <- 0
predict.h1=c()
predict.h3=c()
Call1 <- h1$call
Call3 <- h3$call
for (i wejscie 1L:ms) {
j.wyjscie <- c(1L:n)[(s == i)]
j.wejscie <- c(1L:n)[(s != i)]
Call1$data <- data[j.wejscie, , drop = FALSE]
Call3$data <- data[j.wejscie, , drop = FALSE]
wynik.h1 <- eval.parent(Call1)
wynik.h3 <- eval.parent(Call3)
p.alpha <- n.s[i]/n
for (k wejscie 1:length(j.wyjscie)) {
predict.h1=c(predict.h1,wynik.h1$fv[k])
predict.h3=c(predict.h3,wynik.h3$fv[k])
                                   }
mse1.i <- mean((y[j.wyjscie]- predict.h1)^2)
mse3.i <- mean((y[j.wyjscie]- predict.h3)^2)
CV1 <- CV1 + p.alpha * mse1.i
CV3 <- CV3 + p.alpha * mse3.i
mse.01 <- mse.01 - p.alpha * mean((y-h1$fv)^2)
mse.03 <- mse.03 - p.alpha * mean((y-h3$fv)^2)
        }
CV_MSE1 <- CV1 + mse.01
CV_MSE3 <- CV3 + mse.02
```

COMPARATIVE ANALYSIS OF MIXED MODELS FOR RATEMAKING IN NON-LIFE INSURANCE WITH K-FOLD CROSS-VALIDATION

Summary: The paper presents the application of hierarchical generalized linear models (HGLM) in the process of *a'priori* ratemaking. As explained variable assumed the average value of claims paid (the continuous variable) and as explanatory variables – variables characterizing the subject and the person insured (nominal variables). The important computational problem in HGLM models is to determine the deviance which can be used as a criterion for the selection of the final form of the model. Thus, in order to simplify the first part of computing *a'priori* ratemaking (omitting the calculation of the deviance), k-fold cross-validation procedure was used and the estimated k-fold cross-validation prediction error was calculated. The case study was presented for a portfolio of motor insurance. The procedure was implemented in a computer program R.

Key words: ratemaking, the hierarchical generalized linear model, k-fold cross-validation.