PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu nr 312

RESEARCH PAPERS
of Wrocław University of Economics No. 312

Zagadnienia aktuarialne – teoria i praktyka

Redaktor naukowy
Joanna Dębicka

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Dorota Pitulec

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

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

ISSN 1899-3192
ISBN 978-83-7695-315-1

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 312 • 2013
Zagadnienia aktuarialne – teoria i praktyka ISSN 1899-3192

Daniel Sobiecki
Szkoła Główna Handlowa w Warszawie

DWUSTOPNIOWE MODELOWANIE SKŁADKI ZA UBEZPIECZENIE KOMUNIKACYJNE OC

Streszczenie: Przedmiotem referatu jest zagadnienie dwustopniowego modelowania składki czystej za ubezpieczenie komunikacyjne OC – osobno modeluje się części składki związane ze szkodami umiarkowanymi oraz ekstremalnymi. W modelu zgłoszone szkody są dzielone na dwie mniej heterogeniczne grupy, co przekłada się na poprawę dokładności oceny ryzyka znacznie obciążającego finansowy wynik zakładu ubezpieczeń. Empiryczną część pracy stanowi badanie portfela jednego z zakładów ubezpieczeń działających na polskim rynku, obejmujące ponad 500 tys. polis. Referat w dużej mierze opiera się na pracy magisterskiej napisanej w Instytucie Ekonometrii SGH pod kierunkiem naukowym prof. dr hab. Marii Podgórskiej.

Słowa kluczowe: modelowanie składki, GLM, modele zmiennych licznikowych, metoda *Excesses Over Threshold*.

1. Wstęp

Przedmiotem artykułu jest zagadnienie dwustopniowego modelowania składki czystej za ubezpieczenie komunikacyjne OC – osobno modeluje się część składki związaną ze szkodami umiarkowanymi oraz ekstremalnymi. Inspirację do podjęcia tematu pracy stanowił model szkód ekstremalnych, zaproponowany w monografii [Denuit i in. 2007], w którym zgłoszone szkody dzielone są na dwie mniej heterogeniczne grupy, co przekłada się na poprawę dokładności oceny ryzyka znacznie obciążającego finansowy wynik zakładu ubezpieczeń. Empiryczną część pracy stanowi badanie portfela jednego z zakładów ubezpieczeń działających na polskim rynku, obejmujące ponad 500 tys. polis.

Na początku pracy przedstawione są dane statystyczne wykorzystane w badaniu oraz charakterystyka obserwowanej częstości i wysokości szkód. W dalszej kolejności omawiane są podstawy teorii zdarzeń ekstremalnych, szczególnie metoda *Excesses Over Threshold* (EOT) i uogólniony rozkład Pareto. Zaprezentowane są również wyniki zastosowania teorii na danych empirycznych, m.in. do wyboru progu oddzielającego dwie grupy szkód i dopasowanie uogólnionego rozkładu Pareto. Następnie przedstawiony jest finalny model częstości szkód umiarkowanych oraz model wyso-

kości szkód umiarkowanych. W ostatniej części pracy podsumowane są wnioski płynące z analiz wraz z porównaniem wyników z wynikami zawartymi w monografii [Denuit i in. 2007].

W analitycznej części artykułu wykorzystane są trzy pakiety statystyczno-ekonometryczne: R 2.14, Stata 11 i GenStat 12 oraz arkusz kalkulacyjny MS Excel 2007. Program R najlepiej ze wskazanych służy wizualizacji danych na różnych etapach analizy ze względu na dostępność m.in. takich narzędzi, jak programowalne histogramy, wykresy kwantylowe czy rozbudowane wykresy reszt w ramach analizy dopasowania modelu. Stata najszybciej szacuje złożone modele regresji na dużych zbiorach danych. GenStat z kolei wykorzystany jest w analizie szkód ekstremalnych, ponieważ zawiera oprogramowaną metodę EOT wraz z testami zgodności.

2. Opis portfela polis i szkód

Dane wykorzystane w modelowaniu składki czystej stanowią zbiór 516 695 polis zawartych przez osoby fizyczne od 1 stycznia 2008 r. do 31 grudnia 2008 r., z których zgłoszono łącznie 24 698 szkód. Szkody miały miejsce w latach 2008-2009, ale dotyczą tylko wspomnianego portfela polis. Nie ma informacji o szkodach z 2008 r. z polis zawartych w 2007 r. ani informacji o szkodach z 2009 r. z polis zawartych w 2009 roku. Odpowiada to metodzie roku polisowego, która w odróżnieniu od metody roku kalendarzowego i metody roku wypadku zapewnia dokładne dopasowanie kosztów i okresu ekspozycji na ryzyko (McClenahan w: [Teugels, Sundt 2004]). Proces likwidacji dużych szkód może trwać latami, dlatego metoda roku polisowego wymaga oszacowania „ostatecznej" wysokości szkód. W tym celu wykorzystuje się przeważnie historyczne wzorce rozwoju wysokości szkody i np. metodę łańcuchową. Prezentowane dane przedstawiają jednak stan szkód na koniec marca 2012 roku. Rezerwy na szkody zgłoszone z analizowanego portfela (RBNS) były wtedy zerowe. Rezerwy na szkody zaszłe, ale nie zgłoszone (IBNR) nie są uwzględnione. Tabela 1 prezentuje podstawowe charakterystyki empirycznego rozkładu wysokości szkód wyjściowego portfela.

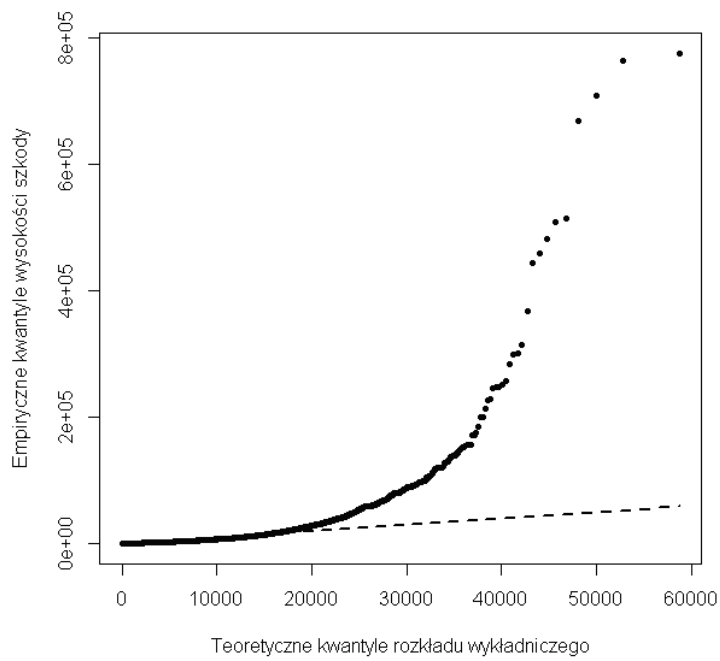
Statystyki zawarte w tab. 1 ukazują duże zróżnicowanie wysokości szkód. Średnia szkoda jest dziewięćdziesięciokrotnie większa od najmniejszej szkody i ponad sto czterdzieści razy mniejsza od szkody największej. Co więcej, górny kwartyl rozkładu jest mniejszy od średniej, czyli ponad 75% szkód jej nie przekracza. Silną prawostronną asymetrię rozkładu potwierdza niska wartość mediany, stanowiącej jedynie 43,7% średniej, oraz wysoka dodatnia wartość współczynnika skośności. Wymienione własności wraz z bardzo wysoką kurtozą świadczą o znacznej różnicy między analizowanym rozkładem empirycznym a rozkładem normalnym, postulowanym w klasycznym modelu liniowym, co implikuje wykorzystanie w dalszej analizie modeli rezygnujących z tego założenia.

Tabela 1. Podstawowe charakterystyki rozkładu wysokości szkód

Minimum	60 zł
Maximum	775 890,88 zł
Średnia szkoda	5 438,47 zł
Odch. stand.	16 556,43 zł
Dolny kwartyl	1 245,35 zł
Mediana	2 378,09 zł
Górny kwartyl	4 838,31 zł
90. percentyl	10 299,65 zł
95. percentyl	17 560,48 zł
99. percentyl	54 810,01 zł
Skośność	22,72
Kurtoza	800,99

Źródło: opracowanie własne.

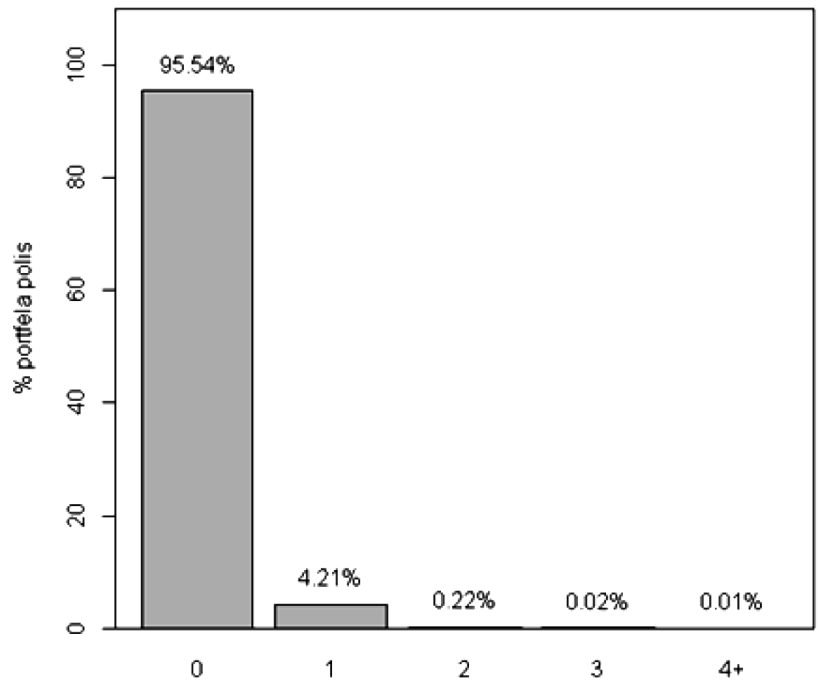
Rozkładem z ciężkim ogonem nazywa się taki rozkład, którego ogon ma większą masę prawdopodobieństwa, niż wynikałoby to z rozkładu wykładniczego [Berlain i in. 2004]. Rysunek 1 przedstawia wykres typu „kwanty-kwantyl" dla rozkładu



Rys. 1. Wykres kwantylowy – rozkład wykładniczy

Źródło: opracowanie własne (R).

wykładniczego. Na osi pionowej umieszczone są kwantyle analizowanego rozkładu wysokości szkody, a na drugiej osi kwantyle porównywanego rozkładu wykładniczego, którego parametry oszacowano MNW na podstawie próby. Badana zmienna zdecydowanie odbiega od zadanego rozkładu teoretycznego, ponieważ wykres nie pokrywa się z przerywaną prostą, która symbolizuje równość kwantyli empirycznych i teoretycznych. Powyżej 20 000 zł kwantyle empiryczne zaczynają rosnąć szybciej od teoretycznych. Przykładowo, maksymalnej obserwowanej wartości szkody odpowiada z rozkładu wykładniczego wartość jedynie 60 000 zł. Wypukłe odchylenia wykresu empirycznych kwantyli wysokości szkody od prostej wskazują zatem na rozkład o ciężkim ogonie.



Rys. 2. Rozkład liczby szkód

Źródło: opracowanie własne (R).

Rysunek 2 pokazuje strukturę portfela ze względu na liczbę szkód. Zdecydowana większość klientów jest bezszkodowa. Częstość szkód mierzona jako stosunek liczby szkód do annualizowanej ekspozycji na ryzyko wynosi 4,78%. W gronie klientów szkodowych 94,4% stanowią klienci, którzy zgłosili dokładnie jedną szkodę. Tylko z jednej na 10 000 polis zgłoszono cztery i więcej szkód. W dalszej analizie punktem wyjścia do modelowania częstości będzie model Poissona, w którym postulowana jest równość średniej i wariancji. Średnia empiryczna częstość szkód

wynosi, jak już wspomniano, 4,78%. Natomiast jej wariancja jest równa 5,3%, co sygnalizuje, że warto sięgnąć po modele uwzględniające większe rozproszenie liczby szkód niż w modelu Poissona.

W analizowanej bazie danych dla każdej polisy dostępna jest informacja o liczbie zgłoszonych szkód, wysokości każdej z nich oraz annualizowanej ekspozycji na ryzyko. Ponadto zawarte są dane o umowie ubezpieczenia, kliencie i pojeździe, które przekładają się na następującą listę zmiennych:

– AGREEMENT_TYPE – przyjmuje wartości tariff (tylko umowa OC) i package (klient wykupił pakiet ubezpieczeń),
– RENEWAL – 0 dla nowych klientów, 1 dla klientów odnawiających umowę,
– PREMIUM_SPLIT – 0 dla płacących składkę w całości, 1 dla dzielących płatność,
– SEX_P – płeć ubezpieczonego, 1 – mężczyzna, 0 – kobieta,
– KIND_OF_DIST – typ miejsca rejestracji auta – country, suburban, urban,
– VOIVODESHIP – województwo,
– CLIENT_AGE – wiek klienta,
– CAR_MAKE – marka auta,
– ENGINE – rodzaj silnika (benzyna/diesel),
– POWER – moc silnika w KM,
– CAPACITY – pojemność silnika w cm^3,
– CAR_AGE – wiek samochodu, 0 dla aut wyprodukowanych w roku początku polisy,
– CO-OWNER – 1 – jeśli jest współwłaściciel, 0 w przypadku braku.

Niektóre zmienne w zbiorze są binarne (np. płeć) lub nominalne o większej liczbie kategorii (np. województwo), ale są i takie, które przyjmują wartości ze znacznie liczniejszego zbioru. Są to: wiek klienta, moc silnika, pojemność i wiek samochodu, których wartości zostały pogrupowane w przedziały. Dyskretyzacja zmiennych nie jest opisywana w tym artykule. Jej wyniki są wykorzystane w dalszym modelowaniu.

3. Zastosowanie teorii wartości ekstremalnych i modele szkód umiarkowanych

3.1. Wybrane aspekty teorii zdarzeń ekstremalnych

Duże szkody zdecydowanie wpływają na wynik finansowy zakładu ubezpieczeń. Przyczyną, dla której uzasadnione są oddzielne analizy szkód dużych (ekstremalnych) oraz pozostałych – umownie zwanych małymi lub umiarkowanymi – jest fakt, że jeden standardowy model dla wszystkich szkód może nie zapewnić wystarczająco dobrego dopasowania rozkładu teoretycznego do danych. Głównym celem takiej analizy jest określenie odpowiedniego progu oddzielającego obydwa rodzaje szkód. Problem ten można rozwiązać, korzystając z teorii zdarzeń (wartości) ekstremalnych

(*Extreme Value Theory*, EVT) i uogólnionego rozkładu Pareto (*Generalized Pareto Distribiution*, GPD), których zastosowanie można znaleźć w pracy [Cebrian, Denuit, Lambert 2003]. Przedstawione tu rozwiązania teoretyczne pochodzą z monografii [Denuit i in. 2007]. Szczegółowy opis teorii zdarzeń ekstremalnych można znaleźć w pracy [Berlain i in. 2004].

Całkowita wartość szkód *i*-tego ubezpieczonego, $i = 1,2,...,n$, w ubezpieczeniu OC może być przedstawiona jako

$$S_i = \sum_{k=1}^{N_i^m} C_{ik} + \sum_{k=1}^{N_i^d} D_{ik}, \tag{1}$$

gdzie: N_i^m jest liczbą małych szkód zgłoszonych przez *i*-tego ubezpieczonego, C_{ik} jest wysokością *k*-tej małej szkody zgłoszonej przez *i*-tego ubezpieczonego, N_i^d jest liczbą dużych szkód zgłoszonych przez *i*-tego ubezpieczonego, D_{ik} jest wysokością *k*-tej dużej szkody zgłoszonej przez *i*-tego ubezpieczonego.

W analitycznej części pracy każda z przedstawionych tu wielkości jest traktowana jako zmienna losowa i osobno modelowana. Zakłada się, że wszystkie cztery zmienne są wzajemnie niezależne. O zmiennych S_i, $i = 1,2,...,n$, zakłada się, że są niezależne i mają taki sam rozkład ze wspólną średnią μ. Zgodnie z prawem wielkich liczb

$$P(\bar{S}^{(n)} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} S_i = \mu) = 1$$

dlatego składka czysta jest modelowana jako oczekiwana wartość szkód

$$\mathbb{E}[S_i] = \mathbb{E}\left[\sum_{k=1}^{N_i^m} C_{ik}\right] + \mathbb{E}\left[\sum_{k=1}^{N_i^d} D_{ik}\right] = \mathbb{E}[N_i^m] \times \mathbb{E}[C_{ik}] + \mathbb{E}[N_i^d] \times \mathbb{E}[D_{ik}]. \tag{2}$$

Model szkód ekstremalnych zastosowany w niniejszej pracy jest określany w literaturze jako metoda EOT (*Excesses Over Threshold*). Jej istota polega na tym, że mając dany ciąg zmiennych losowych *i.i.d.* $\{X_i\}_{i = 1..n}$, analizuje się ciąg $\{X_i - u | X_i > u\}_{i = 1..n}$, czyli nadwyżki przekraczające pewien próg *u*, przy czym zakłada się rozkład Poissona dla liczby nadwyżek.

Niech F_u oznacza nieznaną, wspólną dystrybuantę rozkładu zmiennych $[X_i - u | X_i > u]$, dla $i = 1,2,...,n$. Zatem F_u przedstawia warunkowy rozkład nadwyżek szkód pod warunkiem, że szkody przekraczają próg *u*. Dla rozkładu F_u z odpowiednio dużym progiem *u* dobre dopasowanie zapewnia (uzasadnienie jest podane w (3)) dwuparametrowy rozkład Pareto (GPD) o dystrybuancie $G_{\xi,\beta}(\cdot)$, zwany uogólnionym rozkładem Pareto. Ta dwuparametrowa rodzina rozkładów jest definiowana jako

$$G_{\xi,\beta}(x) = G_\xi\left(\frac{x}{\beta}\right), \qquad \beta > 0,$$

gdzie $G_\xi(x) = \begin{cases} 1 - (1 + \xi x)^{-1/\xi} & \text{dla } \xi \neq 0 \\ 1 - \exp(-x) & \text{dla } \xi = 0, \end{cases}$

dla $x \geq 0$ gdy $\xi \geq 0$ oraz $x \in \left[0, -\frac{1}{\xi}\right]$ gdy $\xi < 0$, przy czym parametr ξ nazywa się indeksem Pareto.

Dla odpowiedniej funkcji $\beta(u)$ i ustalonego ξ, oszacowanych na podstawie danych, przybliżenie

$$F_u(x) \approx G_{\xi,\beta}(x) \text{ dla } x \geq 0 \tag{3}$$

jest dobre dla dużego u. Powyższą aproksymację uzasadnia się formułą zwaną twierdzeniem Pickandsa–Balkemy–de Haana:

$$\lim_{u\to\infty} \sup_{x\geq 0} \left|F_u(x) - G_{\xi,\beta(u)}(x)\right| = 0.$$

Z (3) wynika, że ciąg $\{X_i - u | X_i > u\}_{i=1..n}$ może być traktowany jak próbka losowa z uogólnionego rozkładu Pareto, jeśli u jest wystarczająco duże. Odpowiadając na pytanie, co oznacza „wystarczająco wysoki próg", należy wziąć pod uwagę dwa czynniki:

– zbyt mała wartość u oznacza, że rozkład GPD zostanie odrzucony przez testy dopasowania, a oszacowania będą obciążone; obciążenie może być znaczne, ponieważ wraz z obniżaniem progu przybywa sporo obserwacji (szkód umiarkowanych jest zdecydowanie więcej niż ekstremalnych),
– zbyt duża wartość u skutkuje małą próbą nadwyżek ponad próg, co przekłada się na małą precyzję oszacowań.

Biorąc pod uwagę te dwa czynniki, należy wybrać najmniejszą wartość progu u, dla której GPD jest akceptowalnym przybliżeniem ogona rozkładu szkód.

W analitycznej części artykułu do wyboru wysokości progu służą testy zgodności, Andersona–Darlinga, Craméra–von Misesa i Watsona, badające dopasowanie rozkładu teoretycznego (w tym przypadku rozkładu GPD) do danych. Wyjściowym testem jest test Craméra–von Misesa, a pozostałe dwa testy są jego modyfikacjami (przypisują większą wagę różnicom w ogonach rozkładu). Brak podstaw do odrzucenia hipotezy o zgodności rozkładów teoretycznego i empirycznego przez wszystkie trzy testy stanowi mocną przesłankę do wyboru rozkładu GPD. Istnieje też kilka metod graficznych, które poprzez analizę wykresu prowadzą do wyboru wysokości progu. Jedną z nich, która jest zastosowana jako analiza wstępna względem testów, jest badanie zachowania empirycznej średniej nadwyżki szkody ponad wartość u. Średnią nadwyżkę szkody definiuje się następująco

$$e(u) = \mathbb{E}[X - u | X > u] = \int_0^{+\infty} \left(1 - F_u(x)\right) dx.$$

Można wykazać, że dla zmiennej losowej X o rozkładzie $G_{\xi,\beta}(x)$ średnia nadwyżka szkody jest liniową funkcją progu u i ma postać

$$e(u) = \frac{\beta}{1-\xi} + \frac{\xi}{1-\xi} u \tag{4}$$

jeśli $\beta + u\xi > 0$. Z kolei empiryczną funkcję średniej nadwyżki szkody można oszacować w oparciu o obserwacje $\{x_1, ..., x_n\}$ jako

$$\hat{e}_n(u) = \frac{\sum_{x_i>u} x_i}{\#\{x_i: x_i > u\}} - u = \frac{\sum_{x_i>u}(x_i - u)}{\#\{x_i: x_i > u\}}, \tag{5}$$

czyli jako sumę nadwyżek ponad próg u podzieloną przez liczbę obserwacji większych od u. Wartość progu, powyżej którego empiryczna funkcja nadwyżki jest w przybliżeniu liniowa, może zostać wykorzystana do podziału szkód na małe i duże. Co więcej, gdy na przedziale wysokości szkód $[u, +\infty)$ dopasuje się linię prostą do $\hat{e}_n(u)$, to korzystając ze współczynnika kierunkowego i wyrazu wolnego oraz zależności (4), można wstępnie oszacować parametry rozkładu GPD, ξ i β.

Po określeniu wysokości progu można przystąpić do modelowania wysokości szkód ekstremalnych z wykorzystaniem GPD. Liczebność próby dużych szkód jest z reguły niewielka, co nie pozwala na modelowanie średniej szkody ekstremalnej z wykorzystaniem zmiennych objaśniających. Logarytm funkcji wiarygodności, którą należy maksymalizować w celu otrzymania ocen parametrów przy założonym progu u, ma postać

$$\ln L(\xi, \beta) = -\ln\beta \cdot \#\{x_i | x_i > u\} - (1 + \frac{1}{\xi}) \sum_{i|x_i>u} \ln\left(1 + \frac{\xi}{\beta}(x_i - u)\right).$$

Mała próba ekstremalnych szkód może mieć odzwierciedlenie w niskiej wiarygodności oszacowań. Denuit i in. [2007] wskazują, że w praktyce firma ubezpieczeniowa powinna zebrać informacje o ekstremalnych szkodach z wszystkich lat i przeprowadzić ich szczegółową analizę. Wysokości szkód w takim przypadku powinny zostać skorygowane m.in. z powodu inflacji. Dla ubezpieczycieli dysponujących małym lub średnim portfelem z pomocą może przyjść reasekurator dysponujący znacznie szerszą wiedzą o szkodach ekstremalnych.

Oszacowania $\hat{\xi}$ i $\hat{\beta}$ służą wyznaczeniu $\mathbb{E}D_{ik}$, czyli wartości oczekiwanej zgłoszonej szkody ekstremalnej, z wykorzystaniem następującej własności GPD dla $\xi < 1$:

$$\mathbb{E}X = u + \frac{\beta}{1-\xi}. \tag{6}$$

Liczba szkód ekstremalnych zgłoszonych przez i-tego ubezpieczonego, $N_i^{duże}$, może być modelowana z wykorzystaniem rozkładu Poissona ($N_i^{duże} \sim Poi(\lambda_i^{duże})$). Jest to zgodne z faktem, że szkody ekstremalne zdarzają się całkowicie losowo. W modelu regresji Poissona objaśniającym oczekiwaną częstość dużych szkód można wprowadzić zmienne objaśniające za pomocą wykładniczej funkcji łączącej. Jednak mała próba polis z dużymi szkodami pozwala z reguły jedynie na oszacowanie regresji wyłącznie ze stałą. Korzystając z faktu, że w większości przypadków ubezpieczony zgłasza 0 lub 1 ekstremalną szkodę, model regresji Poissona można zastąpić regresją logistyczną lub probitową. W takim przypadku modelowana byłaby

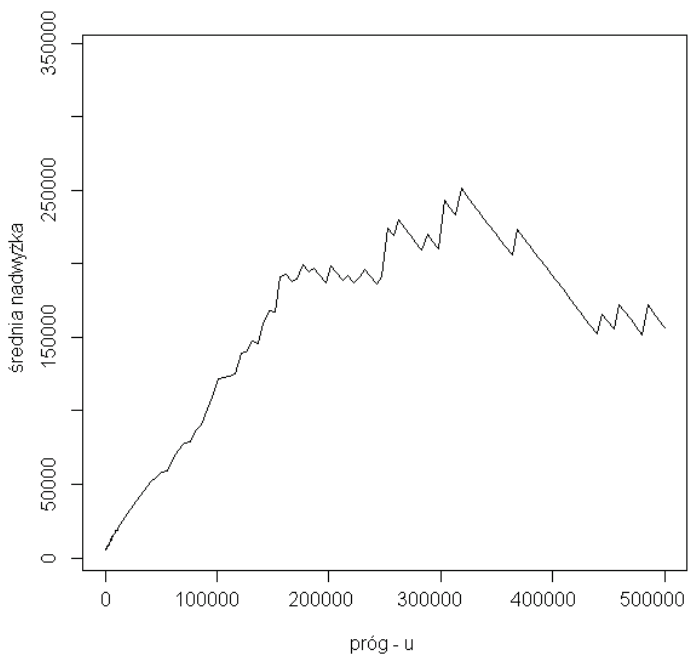
zmienna binarna zamiast zmiennej licznikowej. Obydwa modele zostaną uwzględnione w części analitycznej pracy.

W modelowaniu wysokości szkód umiarkowanych wykorzystane są modele klasy GLM: regresja odwrotna gaussowska, gamma i log-normalna. W modelowaniu częstości tychże szkód zastosowane są modele zmiennych licznikowych – regresja Poissona, ujemna dwumianowa i modele z podwyższoną liczbą zer. Teoria stojącą za tymi modelami nie będzie przybliżana w tym artykule. Można ją poznać m.in. w pracach: [Anderson i in. 2007; Faraway 2006]. Natomiast opis teorii modeli zmiennych licznikowych można znaleźć w [Long, Freese 2001].

3.2. Zastosowanie dwustopniowego modelowania szkód

3.2.1. Wyznaczenie progu i modele szkód ekstremalnych

Próg *u*, powyżej którego szkoda jest klasyfikowana jako ekstremalna i który potrzebny jest do dopasowania uogólnionego rozkładu Pareto, należy wyznaczyć na podstawie testów zgodności jako najmniejszy próg, przy którym nie ma podstaw do odrzucenia hipotezy o zgodności rozkładu empirycznego z teoretycznym. Przed formalnym



Rys. 3. Empiryczna nadwyżka szkody (w zł)

Źródło: opracowanie własne (R).

postępowaniem statystycznym warto jednak skorzystać z jednej z graficznych metod określenia progu. Wybrana została metoda empirycznej średniej nadwyżki szkody ponad próg u.

Rysunek 3 przedstawia przebieg empirycznej średniej nadwyżki szkody ponad próg u dla rozpatrywanego portfela. Zakres zmienności progu został ustalony od 0 do 500 000 zł i jest mniejszy od zakresu zmienności wysokości szkody w portfelu. Wynika to z faktu, że estymatory empirycznej średniej nadwyżki szkody dążą do 0, gdy wysokość progu zbliża się do prawego ogona rozkładu wysokości szkody, ponieważ coraz mniejsza liczba obserwacji przekracza próg (zob. wzór (5)). Zgodnie ze wzorem (4) dla zmiennej losowej X o rozkładzie GPD z dystrybuantą $G_{\xi,\beta}(x)$ średnia nadwyżka szkody jest liniową funkcją progu u. Można przyjąć, że zachowanie empirycznej średniej nadwyżki szkody jest „zachwiane" na końcu przedziału zmienności ze względu na wspomnianą małą liczbę obserwacji. Analiza wykresu pozwala przypuścić, że próg u wynosi ok. 100 000 zł, ponieważ dla u większych od tej wielkości przebieg wykresu średniej nadwyżki może być przybliżony prostą.

Tabela 2 przedstawia wartości statystyk testowych dla progu u = 75 000 zł. Zestawienie informacji z tab. 2 i 3 pokazuje, że dla wszystkich stosowanych testów zgodności nie ma podstaw do odrzucenia hipotezy o zgodności rozkładu szkód ekstremalnych z uogólnionym rozkładem Pareto przy poziomie istotności wynoszącym 5%. Wyjściową wysokością progu było 100 000 zł pochodzące z analizy empirycznej średniej nadwyżki szkody. Później próg był systematycznie obniżany z krokiem 5 000 zł i przy każdej iteracji porównywane były wartości statystyk testowych z wartościami krytycznymi testów zgodności. Próg 75 000 zł był najniższym, dla którego wszystkie trzy testy wskazywały na brak podstaw do odrzucenia hipotezy przy poziomie istotności w wysokości 5%. Dla progu 70 000 zł wszystkie testy wskazują na istotność różnic między empirycznym rozkładem szkód ekstremalnych a teoretycznym rozkładem GPD przy poziomie istotności 5%.

Tabela 2. Wartości statystyk testowych dla progu u = 75 000 zł

Anderson-Darling	0.5920
Cramer-von Mises	0.0868
Watson	0.0762

Źródło: opracowanie własne (GenStat).

Tabela 3. Wartości krytyczne testów zgodności z rozkładem GPD przy poziomie istotności 5%

Anderson-Darling	0.787
Cramer-von Mises	0.126
Watson	0.116

Źródło: opracowanie własne (GenStat).

Tabela 4 przedstawia oszacowania MNW parametrów rozkładu GPD wraz z błędami standardowymi dla progu u wynoszącego 75 000 zł. Błędy są relatywnie niskie, a parametry różnią się istotnie od zera (wskazują na to testy ilorazu wiarygodności).

Tabela 4. Oszacowania parametrów rozkładu GPD (próg u = 75 000 zł)

	Oszacowanie	Błąd stand.
β	35 222	7 739
ξ	0.6312	0.1986

Źródło: opracowanie własne (GenStat).

Próg określający szkody ekstremalne został określony na 75 000 zł, dlatego można już podzielić analizowany portfel na dwie części. Od tego miejsca cały portfel jest nazywany portfelem A, część ze szkodami umiarkowanymi – portfelem B oraz ekstremalnymi – portfelem C. Portfel B jest bardziej homogeniczny od wyjściowego portfela A, zatem jest lepszy do modelowania. Wskazują na to miary rozproszenia, np. odchylenie standardowe zmalało aż o 56%. Nadal mamy do czynienia z rozkładem silnie prawostronnie skośnym, ale współczynnik skośności i kurtoza są w portfelu B o rząd wielkości mniejsze niż w portfelu A.

W portfelu szkód ekstremalnych znalazło się 136 szkód zgłoszonych ze 134 polis ubezpieczenia OC (z dwóch polis zgłoszono po dwie szkody przekraczające próg 75 000 zł). Rozkład wysokości szkód ekstremalnych też jest rozkładem prawostronnie skośnym – współczynnik skośności wynosi 3, a mediana jest mniejsza od średniej.

Próbka 136 szkód ze 134 polis jest rozmiaru, który teoretycznie pozwala na uwzględnienie zmiennych objaśniających w modelach częstości i wysokości szkód ekstremalnych. Jednak empiryczna częstość szkód ekstremalnych wynosi jedynie ok. 0,2‰, dlatego też oba modele będą zawierały tylko wyraz wolny. Odpowiada to założeniu, że szkody ekstremalne są losowe i prawdopodobieństwo ich zgłoszenia jest jednakowe dla każdego klienta zakładu ubezpieczeń (niezależność od wartości charakterystyk). Oczekiwany koszt dużej szkody w takiej sytuacji wyraża się wzorem (6). Po podstawieniu wysokości wybranego progu i oszacowań parametrów otrzymujemy 170 507,05 zł. Empiryczna średnia szkoda ekstremalna wynosi 154 226,4 zł. Jak widać, oszacowanie wielkości oczekiwanej szkody ekstremalnej na podstawie uogólnionego rozkładu Pareto jest tylko o ok. 10% wyższe niż ocena prostego estymatora próbkowego.

Zgodnie z teorią przedstawioną w podpunkcie 3.1 częstość szkód ekstremalnych modelowana jest z wykorzystaniem regresji Poissona z wykładniczą funkcją łączącą i regresji logistycznej. W obydwu modelach nie są uwzględnione zmienne objaśniające, co powoduje ich redukcję do wyrazu wolnego. Oszacowanie MNW tego para-

metru w modelu regresji Poissona wynosi –8,198865, co odpowiada oszacowaniu częstości szkód ekstremalnych na poziomie 0,275‰. Natomiast dla regresji logistycznej mamy odpowiednio oszacowanie wyrazu wolnego wynoszące –8,337721 i częstości 0,2393‰. Oceny częstości uzyskane na podstawie dwóch modeli są względnie bliskie sobie. W dalszym modelowaniu wykorzystana zostanie wyższa wartość częstości z regresji Poissona. Odpowiada to ostrożniejszemu podejściu do ryzyka szkód ekstremalnych.

3.2.2. Modele częstości i wysokości szkód umiarkowanych

Model częstości szkód umiarkowanych w większym stopniu różnicuje taryfę składek niż model wysokości szkód umiarkowanych, ponieważ w procesie estymacji bierze udział większa liczba obserwacji (portfel polis jest zdecydowanie większy niż portfel szkód), a co za tym idzie – można uwzględnić więcej zmiennych objaśniających. W artykule przedstawiona jest jedynie postać ostateczna obu modeli, która jest wynikiem wykorzystania następujących metod: analizy wariancji, testów statystycznych Walda istotności parametrów, usuwania zmiennych nieistotnych, grupowania poziomów zmiennych istotnych, testu ilorazu wiarygodności na występowanie zjawiska większej wariancji od średniej częstości szkód, diagnostyki reszt modeli i kryteriów informacyjnych.

Tabela 5 przedstawia ostateczny model regresji ujemnej dwumianowej dla częstości szkód umiarkowanych. Po zamianie poziomów bazowych i grupowaniu zmiennych prawie wszystkie parametry modelu istotnie różnią się od zera na poziomie istotności 5%. Tylko dla dwóch zmiennych, CLIENT_AGE5 i CAR_MAKE_T6 (nazwa zmiennej uległa zmianie ze względu na grupowanie), empiryczny poziom istotności parametrów nieznacznie przekracza wartość 5%, dlatego też pozostają one w modelu. W przeciwnym wypadku trzeba by zastosować kolejne grupowanie poziomów zmiennych. Ze względu na brak istotności odrzucone zostały dwie zmienne: informacja o rodzaju silnika (benzyna/diesel) i współwłasności.

Model regresji ujemnej dwumianowej jest uogólnieniem modelu regresji Poissona ze względu na większą wariancję zmiennej objaśnianej od jej wartości oczekiwanej. Za zwiększoną wariancję odpowiada parametr α, którego oszacowanie znajduje się na dole tab. 5. Są tam też wyniki testu ilorazu wiarygodności, które wskazują na odrzucenie hipotezy na rzecz hipotezy alternatywnej mówiącej o tym, że model regresji ujemnej dwumianowej jest istotnie lepszy od modelu Poissona.

W analizie uwzględniony był również model uogólniający regresję ujemną dwumianową ze względu na podwyższoną liczbę zer – model ZINB. Jednak to rozszerzenie zostało odrzucone na podstawie testów statystycznych.

Tabela 6 przedstawia oszacowania parametrów w finalnym modelu regresji log-normalnej dla wysokości szkód umiarkowanych. Wszystkie parametry modelu różnią się istotnie od zera na poziomie istotności 5%. Siedem zmiennych okazało się statystycznie istotnych, czyli zgodnie z oczekiwaniami mniej niż w przypadku mo-

Tabela 5. Finalny model częstości szkód umiarkowanych

CLAIM_COUNT	Coef.	Std. Err.	z	P>|z|
AGREEMENT1	.1538298	.0213251	7.21	0.000
RENEWAL	–.1608341	.0142224	–11.31	0.000
PREMIUM_SP~T	.1345873	.0140454	9.58	0.000
SEX_P	–.1456018	.0155727	–9.35	0.000
DIST2	.1566432	.0191835	8.17	0.000
DIST3	.4396419	.0168992	26.02	0.000
VOIVOD_T2	.1977572	.0309643	6.39	0.000
VOIVOD_T3	.2267418	.0329011	6.89	0.000
VOIVOD_T4	.1596667	.0505637	3.16	0.002
VOIVOD_T5	.2282233	.023994	9.51	0.000
VOIVOD_T6	.1344502	.0251377	5.35	0.000
VOIVOD_T7	.1390526	.0269153	5.17	0.000
VOIVOD_T8	.1949061	.0432378	4.51	0.000
VOIVOD_T9	.2090536	.0236638	8.83	0.000
VOIVOD_T10	.1401496	.0325256	4.31	0.000
VOIVOD_T11	.1787906	.0273626	6.53	0.000
VOIVOD_T12	.2210352	.039978	5.53	0.000
CLIENT_AGE1	.6777378	.0566882	11.96	0.000
CLIENT_AGE2	.2802856	.0446484	6.28	0.000
CLIENT_AGE4	.0896001	.0149025	6.01	0.000
CLIENT_AGE5	–.0394704	.0207713	–1.90	0.000
CLIENT_AGE6	.3046504	.0590219	5.16	0.057
CAR_MAKE_T1	.1111246	.0358501	3.10	0.000
CAR_MAKE_T3	.1832418	.0502333	3.65	0.002
CAR_MAKE_T4	–.036952	.0170072	–2.17	0.000
CAR_MAKE_T5	–.1496184	.0561436	–2.66	0.030
CAR_MAKE_T6	.0370995	.0202021	1.84	0.008
CAR_MAKE_T7	–.1636988	.0709212	–2.31	0.066
CAR_MAKE_T8	–.0869471	.0347072	–2.51	0.021
POWER1	.1443166	.029456	4.90	0.012
POWER3	.106138	.0168868	6.29	0.000
CAPACITY1	.2429226	.0979119	2.48	0.000
CAPACITY3	.206874	.0354346	5.84	0.013
CAR_AGE1	–.1157609	.0337253	–3.43	0.000
CAR_AGE3	–.1114261	.0251344	–4.43	0.001
_cons	–3.552875	.0384164	–92.48	0.000
EXPOSURE	(exposure)			
/lnalpha	.428295	.0426728		
alpha	1.534639	.0654874		
Likelihood-ratio test for alpha=0:				
Chibar2(01) = 1101.87 Prob>chibar2 = 0.000				

Źródło: opracowanie własne (Stata).

delu częstości. Dwie zmienne wymagały grupowania poziomów – województwo rejestracji i wiek klienta. Ten model jest najoszczędniejszy z analizowanej triady pod względem liczby parametrów – ma ich czternaście plus wyraz wolny.

Tabela 6. Model finalny regresji log-normalnej dla wysokości szkód umiarkowanych

CLAIM_AMOU~2	Coef.	Std. Err.	z	P>|z|
RENEWAL	–.062093	.01443	–4.30	0.000
PREMIUM_SP~T	.0258806	.0141351	1.83	0.067
SEX_P	.0708229	.0158054	4.48	0.000
DIST1	.0776252	.0166441	4.66	0.000
DIST2	.0991543	.0178226	5.56	0.000
VOIVO_LN2	–.1271037	.0326122	–3.90	0.000
VOIVO_LN3	–.1547945	.0441871	–3.50	0.000
VOIVO_LN4	–.1217673	.0363701	–3.35	0.001
VOIVO_LN5	–.154111	.0325238	–4.74	0.000
VOIVO_LN6	–.1282316	.0402726	–3.18	0.001
CLIENT_AG~N2	.1429092	.045355	3.15	0.002
CLIENT_AG~N3	.0441489	.0143877	3.07	0.002
POWER2	–.1002078	.0280456	–3.57	0.000
POWER3	–.1094879	.0258425	–4.24	0.000
CO-OWNER	–.0448327	.0182034	–2.46	0.014
_cons	7.892407	.0294167	268.30	0.000

Źródło: opracowanie własne (Stata).

Zarówno kryteria informacyjne, jak i diagnostyka reszt modelu wskazują na wybór modelu regresji log-normalnej jako modelu najlepiej opisującego kształtowanie się wysokości szkód umiarkowanych w portfelu B.

4. Składka czysta i dalsza analiza

Do modelowania częstości szkód umiarkowanych wybrano regresję ujemną dwumianową, a do modelowania wysokości tych szkód wykorzystano regresję log-normalną. Natomiast dla szkód ekstremalnych mamy odpowiednio regresję Poissona i uogólniony rozkład Pareto. W przypadku wyboru rozkładu log-normalnego do modelowania wysokości szkód umiarkowanych i regresji ujemnej dwumianowej do modelowania ich częstości wartość składki czystej za szkody umiarkowane można dla *i*-tego klienta obliczyć z wykorzystaniem wzoru

$$\begin{aligned} &\mathbb{E}\left[\sum_{k=1}^{N_i^m} C_{ik}\right] = \mathbb{E}[N_i^m] \times \mathbb{E}[C_{ik}] = \\ &= \omega_i \exp\left(\beta_0^{cz} + \beta_0^{w} + \sum_{j=1}^{p}\left(\beta_j^{cz} + \beta_j^{w}\right)x_{ij} + \frac{\sigma^2}{2}\right). \end{aligned} \tag{7}$$

Jeśli zastosować regresję Poissona do modelowania częstości szkód ekstremalnych i rozkład GPD do modelowania ich wysokości, to składka czysta za szkody ekstremalne wynosi dla każdego klienta

$$\mathbb{E}\left[\sum_{k=1}^{N_i^d} D_{ik}\right] = \mathbb{E}[N_i^d] \times \mathbb{E}[D_{ik}] = \omega_i \exp(\beta_0^{cz}) \left(u + \frac{\beta}{1-\xi}\right). \tag{8}$$

Jeśli pominie się duże szkody, składka czysta dla klasy referencyjnej wyniesie zgodnie ze wzorem (7)

$$\exp\left(\beta_0^{cz} + \beta_0^{w} + \frac{\sigma^2}{2}\right) = \exp\left(-3{,}55 + 7{,}9 + \frac{1.038^2}{2}\right) = 132{,}78 \text{ zł}, \tag{9}$$

gdzie σ^2 jest wariancją rozkładu log-normalnego (oszacowaną w R z wykorzystaniem pakietu GAMLSS). Wielkość ta mówi o tym, jaką składkę czystą zapłaci klient, którego charakterystyki przyjmują wartości bazowe w analizowanych modelach, co odpowiada jednostkowym (100%) mnożnikom składki bazowej w ostatniej kolumnie tab. 6. Ze względu na dwie zmienne: rodzaj miejsca rejestracji i moc silnika, dla których poziom referencyjny jest różny w modelu częstości i modelu wysokości szkód, przez co mnożnik wpływu całkowitego dla żadnego poziomu tych zmiennych nie jest równy 100%, wielkość składki czystej wyznaczona w punkcie (9) nie ma interpretacji.

Skorzystajmy ze wzoru (7), by obliczyć składkę czystą za szkody umiarkowane dla przykładowego klienta. Jego charakterystyki to: mężczyzna, 24 lata, z obszaru podmiejskiego w województwie mazowieckim, samochód marki Toyota z 2004 r., o mocy 75 KM i pojemności 1200 cm^3. Klient wykupuje samo ubezpieczenie OC po raz trzeci w analizowanej firmie (odnawia umowę), płaci składkę w dwóch ratach, a auto nie ma współwłaściciela. Ubezpieczenie wykupione jest na roczny okres ochrony, dlatego ekspozycja na ryzyko wynosi $\omega_i = 1$. Składka czysta za szkody umiarkowane wynosi

$$\begin{gathered}
132.78\,zł \times 100\%\ (korekta\ za\ wykup\ samego\ OC) \\
\times 79.93\%\ (korekta\ za\ odnowienie) \\
\times 114.41\%\ (korekta\ za\ dwie\ raty) \\
\times 92.85\%\ (korekta\ za\ płeć) \\
\times 129.27\%\ (korekta\ za\ obszar\ podmiejski) \\
\times 114.92\%\ (korekta\ za\ województwo\ mazowieckie) \\
\times 153.08\%\ (korekta\ za\ wiek\ z\ przedziału\ 24-27) \\
\times 103.78\%\ (korekta\ za\ markę\ pojazdu - Toyota) \\
\times 99.52\%\ (korekta\ za\ moc\ silnika\ z\ przedziału\ 67-124\ \text{KM}) \\
122.98\%\ (korekta\ za\ pojemność\ silnika\ z\ przedziału\ 901-2500\ \text{cm}^3) \\
\times 100\%\ (korekta\ za\ wiek\ auta\ z\ przedziału\ 1-16) \\
\times 100\%\ (korekta\ za\ brak\ współwłaściciela) \\
= 325.66\ \text{zł}.
\end{gathered}$$

Do tak obliczonej składki czystej należy dodać jednakową dla wszystkich klientów składkę za szkody ekstremalne, która zgodnie ze wzorem (8) wynosi

$$\mathbb{E}[N_i^d] \times \mathbb{E}[D_{ik}] = 0.0275\% \times 170\,507.05 = 46.89\ zł.$$

Zatem składka dla omawianego przykładowego klienta wynosi 372.55 zł.

Tabela 7 przedstawia wybrane wyniki modelowania częstości i wysokości szkód umiarkowanych. Trzecia i czwarta kolumna zawierają odpowiednio $\exp(\beta_j^{freq})$ i $\exp(\beta_j^{cost})$ obliczone dla każdego poziomu wszystkich zmiennych. Łącznie wykorzystano dwanaście z wyjściowych trzynastu zmiennych. Tylko informacja o rodzaju silnika (benzyna/diesel) okazała się nieistotna w wyjaśnianiu zarówno częstości, jak i wysokości szkód umiarkowanych.

Z tab. 7 można odczytać między innymi, że w analizowanym portfelu mężczyzn charakteryzuje niższa częstość szkód niż kobiety, natomiast szkody powodowane przez mężczyzn są wyższe. Klienci odnowieniowi charakteryzują się szkodowością (połączony wpływ częstości i wysokości szkód) o 20% niższą od nowych klientów w portfelu. W mieście częstość szkód jest zdecydowanie wyższa niż na wsi, ale są to

Tabela 7. Wybrane mnożniki składki bazowej

Zmienna	Poziom	Częstość	Wysokość	Wpływ całkowity
Podział płatności	płatność jednorazowa	100%	100%	100,00%
	płatność ratalna	114,41%	102,62%	117,41%
Płeć	kobieta	100%	100%	100,00%
	mężczyzna	86,45%	107,34%	92,79%
Rodzaj miejscowości	wieś	100%	108,07%	108,07%
	teren podmiejski	116,96%	110,42%	129,15%
	miasto	155,22%	100%	155,22%
Moc silnika (w KM)	66-	100%	90,46%	90,46%
	67-124	111,20%	89,63%	99,67%
	125+	115,52%	100%	115,52%
Pojemność silnika (w cm^3)	900-	100%	100%	100,00%
	901-2500	122,98%	100%	122,98%
	2500+	127,50%	100%	127,50%
Wiek samochodu	0	89,07%	100%	89,07%
	1-16	100%	100%	100,00%
	17+	89,46%	100%	89,46%
Współwłasność	jeden właściciel	100%	100%	100,00%
	współwłasność	100%	95,62%	95,62%
Typ umowy	umowa pakietowa	116,63%	100%	116,63%
	umowa OC	100%	100%	100,00%
Odnowienie	nowy klient	100%	100%	100,00%
	odnowienie umowy	85,14%	93,98%	80,02%

Źródło: opracowanie własne.

szkody o mniejszej wysokości. Szkodowość rośnie wraz ze wzrostem takich parametrów samochodu, jak pojemność czy moc silnika. Auta nowe i w wieku powyższej szesnastu lat cechują się częstością niższą od tych z przedziału 1-16 lat o ponad 10%.

5. Wnioski

W pracy tej przedstawiona jest pierwsza w Polsce i, o ile autorowi wiadomo, druga na świecie empiryczna analiza składki przy uwzględnieniu dekompozycji portfela na szkody ekstremalne i umiarkowane z wykorzystaniem nowej metody EOT i rozkładu GPD. W monografii [Denuit i in. 2007] analiza przeprowadzona jest na mniejszym zbiorze danych jednego z belgijskich ubezpieczycieli – ok. 18 tys. szkód zgłoszonych ze 160 tys. polis, przy 8 zmiennych objaśniających, w tej pracy zaś jest odpowiednio ok. 24 tys. szkód zgłoszonych z ponad 500 tys. polis i 12 istotnych zmiennych objaśniających. W obydwu pracach została wybrana regresja ujemna dwumianowa do modelowania częstości szkód umiarkowanych i regresja log-normalna jako model wysokości szkód umiarkowanych. Jednak w tej pracy pod uwagę jest wziętych więcej modeli i dokładniejsza jest procedura wyboru modelu dzięki uwzględnieniu analizy reszt i kryteriów informacyjnych. W monografii [Denuit i in. 2007] częstość szkód ekstremalnych jest oszacowana na poziomie 0,117‰ przy progu 100 tys. euro i maksymalnej szkodzie w wysokości ok. 2 mln euro. W tej analizie częstość szkód ekstremalnych wynosi 0,275‰ przy progu 75 tys. zł i szkodzie maksymalnej w wysokości ok. 775 tys. zł.

Niektóre wnioski są wspólne dla obu analiz. Przykładowo, częstość szkód osób płacących składkę ratalnie jest wyższa od częstości szkód osób opłacających ją jednorazowo. Na wsi jest niższa częstość szkód niż w mieście, ale są to szkody o wyższej wartości. Obie analizy wskazują na to, że częstość szkód rośnie wraz ze wzrostem mocy silnika. Jednak w niniejszym badaniu moc silnika ma wpływ także na wysokość szkód, a u Denuit i in.[2007] jest on nieistotny. Wyniki są zgodne także we wskazaniu na to, że kobiety charakteryzują się wyższą częstością szkód od mężczyzn. Jednak belgijska analiza wskazuje brak wpływu płci klienta na wysokość szkód, a według wyników tej pracy kobiety odpowiadają za szkody o niższej wysokości.

Część wniosków płynących z porównywanych analiz się nie pokrywa. W monografii [Denuit i in. 2007] częstość szkód wśród osób wykupujących jedynie ubezpieczenie OC jest wyższa od częstości szkód tych, którzy wykupują pakiet, a w analizowanym tutaj przypadku jest odwrotnie. W przypadku polskiego portfela wiek ubezpieczanego samochodu nie ma wpływu na wysokość szkód, a w przypadku belgijskim jak najbardziej. Co więcej, tam częstość szkód maleje wraz ze starzeniem się klientów, a tu zależność ma przebieg niemonotoniczny.

Przeprowadzona analiza pozwala wnioskować o wpływie wielu zmiennych na kształtowanie się częstości i wysokości szkód. Jedną z nich jest pojemność silnika.

Częstość szkód rośnie wraz z jej wzrostem. Z kolei klienci odnawiający umowy w analizowanym zakładzie ubezpieczeń mają zarówno niższą częstość, jak i wysokość szkód w porównaniu z osobami ubezpieczającymi się po raz pierwszy. Analiza ze względu na markę pojazdu i województwo prowadzi do wniosku, że najwyższą szkodowością charakteryzują się auta marki BMW w województwie łódzkim, a najniższą auta marki Hyundai w województwie opolskim.

W wyniku analizy ustalone zostały:

1) wysokość składki bazowej za szkody umiarkowane, która wynosi 132,78 zł,

2) wysokość składki za szkody ekstremalne, która dla każdego klienta wynosi 46,89 zł,

3) mnożniki służące kalkulacji składek dla poszczególnych grup klientów.

Jeśli podzieli się portfel na grupy według wartości zmiennych ratingowych, to w analizowanym portfelu największa z nich liczy 401 obserwacji i ma następujące charakterystyki: mężczyzna, w wieku od 28 do 44 lat, z obszaru wiejskiego w województwie podkarpackim, samochód marki Opel w wieku od 1 roku do 16 lat, o mocy od 67 do 124 KM i pojemności w przedziale 901-2500 cm^3, wykup tylko ubezpieczenia OC, nowy klient, składka płacona jednorazowo, jeden właściciel. Składka netto z uwzględnieniem narzutu na szkody ekstremalne wynosi dla takiej kombinacji cech 180,13 zł. Dla porównania najniższa składka czysta w badanym portfelu dla polisy zawartej na rok wynosi 118 zł, a najwyższa 997 zł. Przykładowe charakterystyki klienta z najniższą składką to: mężczyzna, w wieku od 58 do 75 lat, z obszaru wiejskiego w województwie opolskim, samochód marki Fiat w wieku 17 lat lub starszy, o mocy do 66 KM i pojemności do 900 cm^3, wykup tylko ubezpieczenia OC, odnowienie umowy, składka płatna jednorazowo, brak współwłaściciela. Z kolei klient z najwyższą składką w analizowanym portfelu ma następujące cechy: kobieta, w wieku do 23 lat, z miasta w województwie kujawsko-pomorskim, samochód marki Alfa Romeo w wieku od 1 roku do 16 lat, o mocy powyżej 125 KM i pojemności w przedziale 901-2500 cm^3, wykup pakietu ubezpieczeń, nowy klient, płatność składki rozłożona na raty, jeden właściciel.

Nie tylko modele częstości i wysokości szkód dostarczają ważnych wniosków, ale także wstępna analiza danych zawiera ciekawe spostrzeżenia. Przed podziałem szkód na dwie grupy średnia szkoda była wyższa od trzeciego kwartyla rozkładu wysokości szkód. O wysokim zróżnicowaniu odszkodowań świadczy również współczynnik skośności wynoszący aż 22 i kurtoza równa 800. Jeśli chodzi o rozkład liczby szkód, to uwagę zwraca fakt występowania polis, z których zgłoszono więcej niż cztery szkody rocznie. Ponadto są w portfelu dwie polisy, z których zgłoszono po dwie szkody ekstremalne (powyżej 75 000 zł) w badanym roku.

Dotychczasowe rozważania doprowadziły do następujących wniosków ogólnych dotyczących zastosowanych metod:

a) wysokość szkód jest objaśniana przez mniejszą liczbę czynników niż częstość szkód,

b) modele z podwyższoną liczbą zer w analizowanym przypadku nie poprawiają dopasowania regresji Poissona i regresji ujemnej dwumianowej w procesie modelowania częstości szkód,

c) poprawa dokładności klasyfikacji polegająca na zwiększaniu liczby grup jest ograniczona względami praktycznymi (koszty pozyskania dodatkowej informacji, jej weryfikowalność) i istotnością zmiennych,

d) uogólnione modele liniowe i modele zmiennych licznikowych stanowią szeroką klasę modeli z podbudową statystyczną, która pozwala wyciągać ważne dla praktyki ubezpieczeniowej wnioski o kalkulacji składek,

e) teoria wartości ekstremalnych przedstawia szereg metod, dzięki którym można modelować zdarzenia rzadkie, ale znacznie wpływające na finansowy wynik zakładów ubezpieczeń.

Literatura

Anderson D., Feldblum S., Modlin C., Schirmacher D., Schirmacher E., Thandi N., *A Practitioner's Guide to Generalized Linear Models: A Foundation for Theory, Interpretation and Application*, Towers Watson, 2007.

Berlain J., Goegebeur Y., Segers J., Teugels J., de Waal D., Ferro C., *Statistics of Extremes: Theory and Applications*, John Wiley & Sons Ltd., 2004.

Cebrián A.C., Denuit M., Lambert P., *Generalized Pareto fit to the Society of Actuaries' large claims database*, "North American Actuarial Journal" 2003 (7), s. 18-36.

Denuit M., Maréchal X., Pitrebois S., Walhin J.-F., *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-malus Systems*, John Wiley & Sons Ltd., 2007.

Faraway J.J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC, 2006.

Long J.S., Freese J., *Regression Models for Categorical Dependent Variables using Stata*, Stata Press, Texas 2001.

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

Teugels J.L., Sundt B., *Encyclopedia of Actuarial Science*, John Willey & Sons Ltd., 2004.

TWO-STAGE PREMIUM MODELLING IN MTPL

Summary: The subject of this paper is a two-stage modelling of pure premium in MTPL. The parts of premium associated with moderate and extreme claims are modeled separately. In this approach reported claims are divided into two less heterogeneous groups and the model improves the assessment of risk, which poses a significant financial burden for an insurance company. The empirical part of the work presents the study of the portfolio of Polish non-life insurance company, with more than 500,000 policyholders. The paper is largely based on a master's thesis written at the Institute of Econometrics, Warsaw School of Economics under the direction of Prof. Maria Podgórska.

Keywords: premium modelling, GLM, count data models, Excesses Over Threshold method, generalized Pareto distribution.