DYNAMICZNE MODELE EKONOMETRYCZNE
X Ogólnopolskie Seminarium Naukowe, 4–6 września 2007 w Toruniu
Katedra Ekonometrii i Statystyki, Uniwersytet Mikołaja Kopernika w Toruniu

Monika Jeziorska - Pąpka
Uniwersytet Mikołaja Kopernika w Toruniu

Zastosowanie modeli dwumianowych do opisu asymetrii informacji na rynku ubezpieczeń na przykładzie polis komunikacyjnych OC

1.Charakterystyka rynku ubezpieczeń komunikacyjnych

Ubezpieczenie odpowiedzialności cywilnej (zwanej OC) dotyczy każdego posiadacza samochodu. Zakup polisy OC jest obowiązkowy. Klient nabywając taką polisę jest zainteresowany, aby zapłacić jak najmniej, natomiast towarzystwo ubezpieczeniowe zainteresowane jest sprzedażą swoich polis klientom „najlepszym", czyli takim, którzy nie powodują szkód. Problemem dla firmy ubezpieczeniowej jest określenie, przed zwarciem polisy, jaki to rodzaj klienta - „szkodowiec" czy „nieszkodowiec"(zły czy dobry klient).

Ubezpieczyciel może ocenić potencjalnego klienta na podstawie takich cech jak: płeć, miejsce zamieszkania, rodzaj pojazdu, czas posiadania prawa jazdy, liczba szkód komunikacyjnych itp. To pozwala mu zaklasyfikować klienta do odpowiedniej klasy ryzyka. Wszystkie podmioty w jednej klasie płacą jednakową cenę za ubezpieczenie. Oznacza, że klienci mniej skłonni do ryzyka płacą za ubezpieczenie zbyt wygórowaną cenę, nieadekwatną do swojej skłonności do ryzyka i odwrotnie. Takie zaklasyfikowanie do danej grupy jest właśnie wynikiem asymetrii informacyjnej. Im większy stopień asymetrii, tym podmioty w danej grupie są bardziej zróżnicowane.

Asymetria informacji na rynku ubezpieczeń oznacza, że ubezpieczający nie zna indywidualnej skłonności ubezpieczanego do ryzyka. Oznacza to, że decyzje o zakwalifikowaniu do odpowiedniej grupy podejmowane są w oparciu o oszacowane prawdopodobieństwo wystąpienia szkody. Modelami stosowanymi do estymacji prawdopodobieństw są modele dwumianowe (dychotomiczne, binarne).

Celem artykułu jest zaprezentowanie modeli dwumianowych, które mogą być wykorzystywane do określenia, jakie jest oczekiwane prawdopodobieństwo wystąpienia szkody z polisy OC przez potencjalnego klienta.

2. Modele dwumianowe – estymacja i weryfikacja

W modelach dwumianowych przedmiotem wyjaśniania jest prawdopodobieństwo P_i przyjmowania przez zmienną Y jednej z dwóch możliwości (y=1 lub y=0), przy czym jest to prawdopodobieństwo warunkowe, to znaczy - pod warunkiem, że zmienne objaśniające będą kształtować się na określonym poziomie. Prawdopodobieństwo jest funkcją wektora zmiennych objaśniających oraz wektora parametrów, co zapisać można jako:

$$P_i = P(y_i = 1) = F(x_i^T \beta) \qquad i=1,2,...n \tag{1}$$

x_i - wektor zmiennych objaśniających, β - wektor ocen parametrów, $F(\cdot)$ jest funkcją, która przekształca prawdopodobieństwo z przedziału (0, 1) na cały zbiór liczb rzeczywistych.

Ze względu na postać funkcji F wyróżnia się wiele rodzajów modeli dwumianowych. Najczęściej stosowane są modele logitowe i probitowe.

W modelu logitowym funkcja transformacji jest dystrybuantą rozkładu logistycznego:

$$P_i = F(x_i^T \beta) = \frac{1}{1+\exp(-x_i^T \beta)} = \frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)}, \tag{2}$$

$$LP_i = \ln \frac{P_i}{1-P_i} = x_i^T \beta, \tag{3}$$

gdzie*:* LP_i jest logarytmem ilorazu szans zajścia i niezajścia zdarzenia, zwany logitem.

W modelu probitowym przyjmuje się, że prawdopodobieństwa P_i są wartościami dystrybuanty rozkładu normalnego (0,1) w punktach $x_i^T \beta$:

$$P_i = F(x_i^T \beta) = \int_{-\infty}^{\mathbf{x_i \beta}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt . \tag{4}$$

Wartości funkcji odwrotnej do F:

$$x_i^T \beta = F^{-1}(P_i) = G(P_i) \tag{5}$$

nazywa się probitami[1].

[1] W literaturze spotyka się również określenie *normit* , natomiast probitem nazywa się wyrażenie $x_i^T \beta = F^{-1}(P_i) + 5$.

Przekształcenia probitowe stosuje się, gdy prawdopodobieństwa dla zmiennej objaśnianej mają rozkład normalny, bądź zbliżony do normalnego.

Pomiędzy parametrami β w modelu logitowym i probitowym zachodzi następująca zależność[2]: $\beta_{\text{logit}} \approx 1.6\, \beta_{\text{probit}}$.

Parametry modelu dwumianowego dla zbioru mikrodanych szacuje się metodą największej wiarygodności. Logarytm funkcji wiarygodności wynosi:

$$\ln L = \sum_{i=1}^{n} \left\{ y_i \ln F\left(x_i^T \beta\right) + \left(1 - y_i\right) \ln\left[1 - F\left(x_i^T \beta\right)\right] \right\}. \tag{6}$$

Jako miary dopasowania stosuje się współczynniki determinacji R^2 Efrona (oparty na teoretycznych wartościach prawdopodobieństwa) oraz R^2 McFaddena (oparty na wartości funkcji wiarygodności)[3]. W modelach dla dużych zbiorów mikrodanych niska wartość współczynnika determinacji jest typowa i nie oznacza jego nieistotności.

Model dwumianowy pozwala ustalić zarówno prognozę prawdopodobieństwa jak i prognozę zmiennej *y*. Przekształcenie prawdopodobieństwa na zmienną dychotomiczną odbywa się według *standardowej zasady prognozy*:

$$\hat{y} = 1 \quad \text{jeśli} \quad \hat{P}_i > 0.5 \quad \text{oraz} \quad \hat{y} = 0 \quad \text{jeśli} \quad \hat{P}_i \le 0.5$$

W sytuacji, gdy mamy próbę niezbilansowaną, tzn. taka, w której liczba jedynek znacznie różni od ilości zer do prognozowania należy zastosować modyfikację standardowej zasady i liczyć prognozy według *zasady optymalnej wartości granicznej* α, tj:

$$\hat{y} = 1 \quad \text{jeśli} \quad \hat{P}_i > \alpha \quad \text{oraz} \quad \hat{y} = 0 \quad \text{jeśli} \quad \hat{P}_i \le \alpha$$

Wartość graniczną α ustala się jako udział jedynek w próbie.

Po przekształceniu prawdopodobieństwa na *y*=0 i *y*=1 można ocenić jakość prognoz korzystając z tablicy trafień (Tabela 1).

Model dobrze sprawdza się w prognozowaniu, gdy IT > 1 oraz zliczeniowy R^2 >50%. Oznacza to, że klasyfikacja na podstawie modelu jest lepsza od przypadkowej.

[2] W praktyce często zamiast wartości 1.6 stosuje się wartość 1.7.

[3] Pozostałe miary dopasowania szerzej zostały opisane w Gruszczyński (2001).

Tabela 1. Tablica poprawności klasyfikacji

Rzeczywiste	**Przewidywane**		**Razem**	**Trafność**	
	Y=0	Y=1		**Iloraz trafień (IT)**	$\frac{n_{00} \cdot n_{11}}{n_{01} \cdot n_{10}}$
Y=0	n_{00}	n_{01}	$n_{0.}$	**Trafność łączna (zliczeniowy R^2)**	$\frac{n_{00} + n_{11}}{n}$
Y=1	n_{10}	n_{11}	$n_{1.}$	**Trafność Y=0**	$\frac{n_{00}}{n_{0.}}$
Razem	$n_{.0}$	$n_{.1}$	n	**Trafność Y=1**	$\frac{n_{11}}{n_{1.}}$

Źródło: Gruszczyński (2001).

3.Modele dwumianowe – analiza empiryczna

Do modelowania prawdopodobieństw wykorzystano dane z jednej z firm ubezpieczeniowych działających w Polsce. Próba empiryczna obejmowała 506 rocznych polis OC zawartych 2005 roku. Polisy dotyczyły samochodów osobowych. Zbiór potencjalnych zmiennych objaśniających został ustalony na podstawie wniosków ubezpieczeniowych. W skład tych zmiennych zaliczono 2 zmienne ciągłe (wiek właściciela samochodu, lata eksploatacji pojazdu) oraz 4 zmienne jakościowe (płeć, miejsce zamieszkania, marka samochodu oraz pojemność silnika). Wszystkie polisy dotyczyły osób, które posiadają 60% zniżek za bezszkodowość. *Y* – oznacza wystąpienia szkody na danej polisie. Dla zmiennych jakościowych utworzono odpowiednie zmienne binarne.

Na podstawie zmodyfikowanej macierzy korelacji[4] w modelu uwzględniono zmienne skorelowane z Y, a następnie dokonano eliminacji nieistotnych zmiennych metodą a posteriori. Zmienną „wiek właściciela pojazdu" rozpatrywano w wersji zmiennej ciągłej, jak również dzieląc ją na 3 kategorie. Obie wersje tej zmiennej okazały się nieskorelowane z ze zmienną *Y*.

Oszacowano modele logitowy i probitowy. Dla obu modeli otrzymano jednakowy zestaw zmiennych objaśniających. Ponieważ próba była niezbilansowana prognozę *Y* policzono według zasady optymalnej wartości granicznej dla α=0.235 (119/506).

[4] Zob. Gruszczyński (2001)

Tabela.2. Zbiór potencjalnych zmiennych objaśniających

Zmienna	Opis zmiennej (kategorie)		Liczebność
Y	Szkoda	1-tak	119
G1	Płeć	1-mężczyzna	393
C1	Miejsce zamieszkania	Miasta powyżej 50 tys. mieszkańców	220
C2		Miejscowości od 10 do 50 tys. mieszkańców	110
C3		Miejscowości poniżej 10 tys. mieszkańców	176
W1	Wiek właściciela samochodu	do 35 lat włącznie	160
W2		od 36 do 50	196
W3		powyżej 50	150
M1	Marka samochodu	Audi, BMW, Mercedes, Chrysler, Rover, Volvo, VW	66
M2		Citroen, Renault, Peugeot, Lancia	77
M3		Daewoo, Hundai, Kia, Fiat, FSO	87
M4		Ford, Opel, Seat, Skoda	189
M5		Honda, Mazda, Nissan, Suzuki, Toyota, Mitsubishi	87
L	Lata eksploatacji samochodu – zmienna ciągła		
L1	Lata eksploatacji samochodu	Do 3 lat włącznie	192
L2		od 3 do 10 lat	144
L3		Powyżej 10 lat	170
P1	Pojemność silnika (w cm^3)	do 1300	58
P2		1300-1500	120
P3		1500-1700	111
P4		1700-1900	92
P5		powyżej 1900	125

Źródło: opracowanie własne.

Tabela 3. Wyniki estymacji modelu logitowego

Zmienna	Ocena parametru	Błąd stand.	t-Student	Wartość p	χ^2 Walda	p-value	Iloraz szans
Stała	-1.05519	0.26058	-4.04938	0.00006	16.39749	0.00005	2.87252
C1	0.65090	0.23193	2.80646	0.00520	7.87620	0.00501	0.52158
M3	-0.82922	0.33548	-2.47170	0.01378	6.10932	0.01345	2.29153
L	0.04157	0.02131	1.95101	0.05161	3.80644	0.05106	1.04245
Średnia dla Y				0.235			
R^2 McFaddena				4.54%			
R^2 Efrona				4.65%			
Logarytm wiarygodności				-263.476			
Test ilorazu wiarygodności				χ^2=25.05 (p=0.000015)			
Kryterium Akaika				534.951			

Źródło: obliczenia własne.

Tabela 4. Poprawność klasyfikacji klientów firmy ubezpieczeniowej według modelu logitowego

Rzeczywiste	Przewidywane		Razem	Trafność	
	Y=0	*Y*=1		Iloraz trafień (IT)	2.52
Y=0	247	140	387	Trafność łączna	62.65%
Y=1	49	70	119	Trafność *Y*=0	63.82%
Razem	296	210	506	Trafność *Y*=1	58.82%

Źródło: obliczenia własne.

Tabela 5. Wyniki estymacji modelu probitowego

Zmienna	Ocena parametru	Błąd stand.	t-Studenta	Wartość p
stała	-0.64531	0.14902	-4.33046	0.00002
C1	0.38543	0.13420	2.87203	0.00425
M3	-0.48328	0.18550	-2.60521	0.00945
L	-0.02415	0.01197	-2.01712	0.04422
Średnia dla Y	0.235			
R^2 McFaddena	4.59%			
R^2 Efrona	4.688%			
Logarytm wiarygodności	-263.342			
Test ilorazu wiarygodności	χ^2=25.319 (p=0.000013)			
Kryterium Akaika	534.984			

Źródło: obliczenia własne.

Tabela 6. Poprawność klasyfikacji klientów firmy ubezpieczeniowej według modelu probitowego

Rzeczywiste	Przewidywane		Razem	Trafność	
	Y=0	*Y*=1		Iloraz trafień (IT)	100.82
Y=0	247	70	317	Trafność łączna	76.48%
Y=1	49	140	189	Trafność *Y*=0	77.92%
Razem	296	210	506	Trafność *Y*=1	74.07%

Źródło: obliczenia własne.

Na podstawie powyższych modeli można powiedzieć, że mieszkańcy miejscowości powyżej 50 tys. mieszkańców (zmienna C1) mają większe szanse spowodowania szkody, natomiast posiadacze pojazdów marki: Daewoo, Hundai, Kia, Fiat, FSO (kategoria M3) maja mniejsza skłonność do szkody niż właściciele innych pojazdów. Oceny parametrów w modelu logitowym i probitowym spełniają zależność $\beta_{\text{logit}} \approx 1.7\,\beta_{\text{probit}}$, przy czym znak przy zmiennej L jest różny w każdym z modeli. Stanowi to pewną rozbieżność w interpretacji. Na podstawie modelu logitowego stwierdza się, że im starsze auto tym większe prawdopodobieństwo spowodowania szkody. Natomiast według modelu probitowego zależność ta kształtuje się odwrotnie.

Aby rozwikłać tę rozbieżność oszacowano model logitowy i probitowy, w których zmienną L podzielono na 3 kategorie. Po eliminacji a posteriori otrzymano modele, w których zmienne C1, M3 i L1 okazały się istotne statystycznie.

Tabela 7. Wyniki estymacji modelu logitowego – zmienne binarne

Zmienna	Ocena parametru	Błąd stand.	t-Student	Wartość p	χ^2 Walda	Wartość p	Iloraz szans
Stała	-1.57118	0.17206	-9.13183	0.00000	83.39037	0.00000	0.20780
C1	0.65186	0.23287	2.79928	0.00532	7.83596	0.00512	1.91910
M3	-0.81174	0.33551	-2.41943	0.01590	5.85364	0.01555	0.44408
L1	0.45308	0.23330	1.94206	0.05269	3.77160	0.05214	1.57315
Średnia dla Y 0.235							
R^2 McFaddena 4.58%							
R^2 Efrona 4.572%							
Logarytm wiarygodności 263.574							
Test ilorazu wiarygodności χ^2=24.855 (p=0.00002)							
Kryterium Akaika 534.72							

Źródło: obliczenia własne.

Tabela 8. Poprawność klasyfikacji klientów firmy ubezpieczeniowej według modelu logitowego – zmienne binarne

Rzeczywiste	Przewidywane		Razem	Trafność	
	Y=0	*Y*=1		Iloraz trafień (IT)	2.579
Y=0	233	154	387	Trafność łączna	60.87%
Y=1	44	75	119	Trafność *Y*=0	60.21%
Razem	277	229	506	Trafność *Y*=1	63.03%

Źródło: obliczenia własne.

Tabela 9. Wyniki estymacji modelu probitowego – zmienne binarne

Zmienna	Ocena parametru	Błąd stand.	t-Student	Wartość p
stała	-0.94788	0.09677	-9.79506	0.00000
C1	0.38104	0.13537	2.81482	0.00507
M3	-0.47354	0.18554	-2.55222	0.01100
L1	0.27476	0.13702	2.00524	0.04547
Średnia dla Y 0.235				
R^2 McFaddena 4.64%				
R^2 Efrona 4.623%				
Logarytm wiarygodności 263.3894				
Test ilorazu wiarygodności χ^2=25.207 (p=0.00001)				
Kryterium Akaika 534.386				

Źródło: obliczenia własne.

Tabela 10. Poprawność klasyfikacji klientów firmy ubezpieczeniowej według modelu probitowego – zmienne binarne

Rzeczywiste	Przewidywane		Razem	Trafność	
	Y=0	*Y*=1		Iloraz trafień (IT)	10.873
Y=0	233	75	308	Trafność łączna	76.48%
Y=1	44	154	198	Trafność *Y*=0	75.65%
Razem	277	229	506	Trafność *Y*=1	77.78%

Źródło: obliczenia własne.

Zarówno w modelu logitowym i probitowym znaki stojące przy zmiennych są takie same. Oznacza to, że jeżeli samochód ma do 3 lat, to prawdopodobieństwo spowodowania szkody takim samochodem, w porównaniu do samochodów starszych, jest wyższe.

Model probitowy okazał się lepszy niż model logitowy w sensie trafności prognoz, jednak żaden z tych modeli nie był testowany na próbie odłożonej. Niskie wartości współczynników determinacji okazały się istotne.

4. Podsumowanie

Modele dwumianowe mogą być przydatnym narzędziem do opisu rynku ubezpieczeń, gdzie kluczowe znaczenie odgrywa znajomość prawdopodobieństwa spowodowania szkody. Asymetria informacji na tym rynku polega na tym, ze ubezpieczony zna swoją skłonność do spowodowania szkody, a ubezpieczyciel tej skłonności nie zna. Oszacowane na podstawie modelu dwumianowego prawdopodobieństwa mogą posłużyć ubezpieczylowi do wyznaczenia odpowiedniej ceny za polisę OC, adekwatnej do oszacowanego prawdopodobieństwa szkody lub zakwalifikowania danej jednostki do odpowiedniej klasy.
Rozmiar asymetrii informacji lepiej zmierzyć na podstawie polis autocasco, gdzie jest dokładnie określona maksymalna suma ubezpieczenia oraz udziały własne w szkodzie. Dane takie pozwoliłyby również na zbadanie, czy na tym rynku występuje pokusa nadużycia.

Literatura

Goldberger, A.S. (1972), *Teoria ekonometrii*, PWE, Warszawa.

Gruszczyński, M. (2000), Dobór zmiennych objaśniających do modelu logitowego, *Przegląd Statystyczny*. XLVII, 175-185.

Gruszczyński, M. (2001), *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Wydawnictwo SGH, Warszawa.

Maddala, G.S., Nelson, F.D, (1974), Analysis of Qualitative Variables, Working Paper, Nr 70, National Bureau of Economic Research, Cambridge.

Wiśniewski, J.W. (1986), *Ekonometryczne badanie zjawisk jakościowych. Studium metodologiczne*, Uniwersytet Mikołaja Kopernika, Toruń.