PRZEGLĄD STATYSTYCZNY
R. LVIII – ZESZYT 1-2 – 2011

BEATA JACKOWSKA

EFEKTY INTERAKCJI MIĘDZY ZMIENNYMI OBJAŚNIAJĄCYMI W MODELU LOGITOWYM W ANALIZIE ZRÓŻNICOWANIA RYZYKA ZGONU

1. WSTĘP

Modele regresji logistycznej (modele logitowe) wykorzystywane są do objaśniania zmiennych jakościowych w zależności od poziomu zmiennych egzogenicznych (jakościowych bądź ilościowych). Regresja logistyczna znajduje ważne zastosowanie m. in. w modelowaniu ryzyka znalezienia się jednostki badania w pewnym stanie. Jeżeli zmienna objaśniana przyjmuje dwa stany, tzn. mówi, czy badane zjawisko wystąpiło, czy też nie, to mamy do czynienia z modelem dwumianowym.[1]

Współcześnie modele logitowe znalazły powszechne zastosowanie w bankach do oceny ryzyka kredytowego oraz w przedsiębiorstwach do oceny lojalności klientów. Są także jednym z narzędzi wykorzystywanym przez aktuariuszy do oceny ryzyka ubezpieczeniowego oraz oceny szansy konwersji i retencji polis ubezpieczeniowych [7]. W ubezpieczeniach na życie model logitowy pozwala na oszacowanie prawdopodobieństwa śmierci w zależności od podstawowych cech demograficznych, takich jak płeć, wiek, miejsce zamieszkania [9], a w przypadku posiadania odpowiednio dużych baz danych dotyczących historii ubezpieczonych do modelu można włączyć informacje zbierane za pomocą ankiet medycznych dołączanych do wniosku ubezpieczeniowego.

W dziedzinie demografii na ogół poszukuje się parametrycznych (analitycznych) modeli ludzkiego procesu przeżycia (tzw. praw wymieralności[2]) jedynie w zależności od wieku budując oddzielnie modele dla kobiet i mężczyzn (ewentualnie dla osób mieszkających w miastach i na wsi). W tym celu wykorzystuje się regresję krzywoliniową. Jako modele analityczne współczynników zgonu najczęściej są stosowane funkcje: wykładnicze, potęgowe, wielomianowe, wielomianowo-wykładnicze, logistyczne (por. np. [11]). Dostęp do baz danych coraz lepszej jakości oraz rozwój metod numerycznych i oprogramowania komputerowego sprawiają, że modele te są coraz bardziej rozbudowane. W literaturze można odnaleźć wiele prób stworzenia demograficznych modeli analitycznych, lecz rzadziej spotkać można zastosowania uogólnionych modeli

[1] Model wielomianowy jest wykorzystywany, jeżeli zmienna zależna może przyjmować więcej niż dwa niezależne stany – mamy wówczas do czynienia z modelem ryzyka konkurencyjnego, tzn. wszystkie zdarzenia są wzajemnie niezależne i suma prawdopodobieństw ich wystąpienia wynosi 1.

[2] Do pierwszych znanych praw wymieralności należą m. in. prawo de Moivre (1725), prawo Gompertza (1825), prawo Makehama (1860), prawo Weibulla (1939). [11]

liniowych (GLM – *generalised linear models*), w tym regresji logistycznej, do analizy danych demograficznych [7], [9], [10]. Metody regresji logistycznej pozwalają na znalezienie statystycznie istotnych czynników ryzyka zgonu oraz zbadanie efektów interakcji między tymi czynnikami, a dodatkowym atutem modelu logitowego jest możliwość interpretacji jego parametrów.

O ile metody regresji logistycznej są szeroko opisane w literaturze i znajdują coraz więcej zastosowań, to mniej uwagi poświęca się modelowaniu interakcji [6]. Przyjmuje się, że efekt interakcji występuje, jeżeli wpływ zmiennej niezależnej na zmienną zależną zmienia się w zależności od wartości innej zmiennej niezależnej. Analiza efektów interakcji między zmiennymi objaśniającymi w regresji logistycznej została dokonana poprzez wprowadzenie do modelu iloczynu tych zmiennych. Szczególna uwaga została zwrócona na interpretację parametrów modelu, która zależy od sposobu kodowania zmiennych oraz uwzględnienia efektów interakcji między zmiennymi objaśniającymi.

Dopasowanie oraz zdolność predykcyjna modelu zależą od jakości danych oraz liczebności grup jednostek wyodrębnionych według wariantów analizowanych cech. W szczególności dla osób o zaawansowanym wieku ubezpieczyciele w Polsce nie posiadają wystarczająco dużo obserwacji, lecz mogą wykorzystać dane demograficzne gromadzone przez GUS. Estymacja modelu przeżycia dla osób starszych jest przede wszystkim niezbędna przy konstrukcji dobrowolnych ubezpieczeniowych produktów emerytalnych, jak dotąd słabo rozpowszechnionych w Polsce.

W niniejszym artykule dwumianowy model logitowy został wykorzystany do analizy ryzyka zgonu osób starszych (w wieku co najmniej 60 lat) w województwie pomorskim w 2009 roku w zależności od podstawowych cech demograficznych. Celem tej analizy była identyfikacja predyktorów ryzyka zgonu oraz odkrycie efektów interakcji między zmiennymi objaśniającymi.

2. POSTAĆ MODELU LOGITOWEGO

Dwumianowy model regresji logistycznej (model logitowy) wykorzystywany jest do objaśniania dychotomicznej zmiennej jakościowej Y w zależności od poziomu zmiennych egzogenicznych $X_1,\ X_2,\ \ldots,\ X_k$ (jakościowych bądź ilościowych). Zmienna objaśniana reprezentowana jest zwykle przez zmienną zero-jedynkową:

$$Y = \begin{cases} 1 & \text{zdarzenie wystąpiło} \\ 0 & \text{zdarzenie nie wystąpiło} \end{cases}. \tag{1}$$

Model logitowy jest szczególnym przypadkiem uogólnionego modelu liniowego [10]:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k, \tag{2}$$

gdzie β_0 jest wyrazem wolnym; $\beta_1,\ \beta_2,\ \ldots,\ \beta_k$ są współczynnikami regresji; g jest funkcją wiążąca (*link function*) określającą związek średniej wartości zmiennej objaśnianej $\mu = E(Y|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$ z liniową kombinacją predyktorów.

W modelu logitowym $\mu = p = P(Y = 1|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$, a funkcja wiążąca nazywana logitem ma postać

$$g(p) = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right). \tag{3}$$

Podsumowując, model logitowy można zapisać w następującej postaci

$$p = P(Y = 1|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{\exp\left(\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right)}. \tag{4}$$

Parametry modelu β_0, β_1, $\ldots$, β_k estymuje się najczęściej metodą największej wiarygodności maksymalizując logarytm funkcji wiarygodności względem parametrów modelu za pomocą iteracyjnych procedur numerycznych[3].

Jeżeli wśród zmiennych objaśniających znajdują się zmienne jakościowe, to wprowadza się je do modelu odpowiednio kodując (*dummy coding*, *indicator coding*) [1]. Zwykle, gdy zmienna ma m wariantów, to wprowadza się $m - 1$ zmiennych zero-jedynkowych (*dummy variables*). Grupa jednostek badania, dla której wartości wszystkich zmiennych objaśniających są równe zero nazywa się grupą referencyjną (*reference group*). Badacz kodując predyktory za pomocą zmiennych zero-jedynkowych ustala arbitralnie grupę referencyjną (np. wybierając grupę najliczniejszą, największego lub najmniejszego ryzyka), która stanie się grupą odniesienia przy interpretacji parametrów modelu.[4]

Zaletą modelu logitowego jest możliwość interpretacji parametrów e^{β_i}. W tym celu wykorzystuje się pojęcie szansy (*odds*) definiowanej jako iloraz prawdopodobieństwa wystąpienia zdarzenia oraz prawdopodobieństwa nie wystąpienia zdarzenia. W rozważanym modelu (4) szansę można wyrazić jako funkcję zmiennych objaśniających:

$$\frac{p}{1-p} = \gamma(x_1, x_2, \ldots, x_k) = \exp\left(\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right). \tag{5}$$

W przypadku wyrazu wolnego, wartość e^{β_0} jest interpretowana jako szansa wystąpienia zjawiska w grupie referencyjnej.

[3] W pracy [3] przedstawiono metody estymacji w przypadku makrodanych, tzn. gdy „(...) w miejsce obserwacji 0-1 dla zmiennej Y mamy dane jedynie frakcje tych obserwacji w grupach jednostek, których indywidualne cechy nie są rozróżnialne" [3, s. 87].

[4] Dyskusję o innych sposobach kodowania można znaleźć w [5] i [8]. Sposób kodowania ma wpływ na wartość współczynników regresji oraz na interpretację parametrów modelu (zmienia się grupa referencyjna), natomiast nie wpływa na wartość prognozowanego prawdopodobieństwa.

Wpływ przyrostu wartości zmiennych niezależnych o Δx_i ($i = 1, 2, \ldots, k$) na szansę wystąpienia zjawiska można określić wyznaczając iloraz szans (*odds ratio*):

$$\psi(x_1, x_2, \ldots, x_k; \Delta x_1, \Delta x_2, \ldots, \Delta x_k) = \frac{\gamma(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_k + \Delta x_k)}{\gamma(x_1, x_2, \ldots, x_k)} = \exp\left(\sum_{i=1}^{k} \beta_i \Delta x_i\right). \tag{6}$$

Jeżeli X_i ($i = 1, 2, \ldots, k$) jest zmienną zero-jedynkową, to e^{β_i} jest równy ilorazowi szans dla grupy, w której $X_i = 1$ oraz grupy, w której $X_i = 0$, przy pozostałych zmiennych jednakowych. Natomiast, gdy zmienna ta jest zmienną ilościową, to iloraz szans e^{β_i} mówi, jak zmieni się szansa, jeżeli zmienna X_i wzrośnie o 1 jednostkę przy pozostałych zmiennych ustalonych.

3. MODELOWANIE INTERAKCJI W REGRESJI LOGISTYCZNEJ

W literaturze istnieją różne definicje efektu interakcji. Według [6, s. 12]: „Mówimy, że istnieje efekt interakcji, kiedy wpływ zmiennej niezależnej na zmienną zależną różni się w zależności od wartości trzeciej zmiennej nazywanej zmienną moderatorem". W regresji logistycznej moderator jest więc zmienną niezależną, której wartości mają wpływ na siłę i/lub kierunek podstawowej zależności.

W modelu logitowym z dwoma predyktorami

$$\text{logit}(p) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 \tag{7}$$

niech X_2 będzie moderatorem zmiennej X_1. Wpływ zmiennej X_1 na zmienną objaśnianą zależy więc od wartości zmiennej X_2, co oznacza, że parametr α_1 jest funkcją zmiennej X_2. Założenie, że współczynnik przy zmiennej X_1 jest liniową funkcją moderatora (np. [4] i [6])

$$\alpha_1 = \alpha_0' + \alpha_3 X_2 \tag{8}$$

prowadzi do modelu postaci

$$\text{logit}(p) = \alpha_0 + (\alpha_0' + \alpha_3 X_2) X_1 + \alpha_2 X_2 = \alpha_0 + \alpha_0' X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2, \tag{9}$$

w którym pojawia się zmienna interakcyjna będąca iloczynem zmiennych objaśniających. Zmieniając oznaczenia parametrów regresji otrzymujemy ostateczną postać modelu

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2. \tag{10}$$

Zgodnie z założeniem model (10) opisuje wpływ X_1 na Y, gdzie X_2 jest moderatorem, a ponieważ powyższa funkcja jest symetryczna ze względu na zmienne X_1 i X_2, więc model (10) opisuje także wpływ X_2 na Y, gdzie X_1 jest moderatorem. Modelowanie interakcji w regresji logistycznej przy użyciu iloczynu zmiennych nie wymaga identyfikacji, która zmienna niezależna jest moderatorem, a która podlega moderowaniu.

Określenie, która zmienna jest moderatorem jest arbitralne – mówi jedynie o punkcie widzenia badacza i nie ma wpływu na oszacowanie współczynników regresji, tylko na interpretację modelu.

Kontynuując powyższe rozumowanie model z trzema predyktorami uwzględniający wszystkie efekty interakcji (interakcję stopnia trzeciego oraz wszystkie interakcje stopnia drugiego) ma postać

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3. \quad (11)$$

Model zawierający wszystkie możliwe zmienne interakcyjne należy do modeli hierarchicznie dobrze zdefiniowanych (typu HWF – *hierarchically well-formulated*). Jeżeli model hierarchicznie dobrze zdefiniowany zawiera jedynie dychotomiczne zmienne niezależne, to prawdopodobieństwo wystąpienia zdarzenia oszacowane na podstawie modelu jest równe prawdopodobieństwu empirycznemu w grupach jednostek wyodrębnionych według wariantów wszystkich niezależnych zmiennych występujących w modelu. (Por. [3, s. 83])

Konstruując model należy pamiętać, że jeżeli predyktor jakościowy mający $m > 2$ wariantów jest reprezentowany przez $m - 1$ zmiennych zero-jedynkowych, to nie jest dozwolone wprowadzenie do modelu iloczynów tych zmiennych (gdyż oznaczałoby to interakcję między wariantami jednej cechy) [4].

Ocenę poprawy dopasowania modelu poprzez wprowadzenie zmiennej interakcyjnej można dokonać porównując miary dobroci dopasowania (np. miary typu pseudo-R^2) modelu zawierającego wyrażenie reprezentujące interakcję oraz modelu nie zawierającego wyrażenia reprezentującego interakcję. Weryfikację statystycznej istotności zmiennych interakcyjnych przeprowadza się w analogiczny sposób, jak pojedynczych zmiennych objaśniających [6], tzn. przy pomocy:

1. testu ilorazu wiarygodności służącego do sprawdzenia, czy wartość funkcji wiarygodności wzrosła statystycznie istotnie poprzez wprowadzenie do modelu zmiennej interakcyjnej;
2. testu Walda służącego do sprawdzenia istotności współczynnika regresji przy zmiennej interakcyjnej.

Szczególną uwagę należy zwrócić na interpretację parametru e^{β_i} dla iloczynu zmiennych. Dodatkowo obecność iloczynu zmiennych w modelu powoduje także zmianę interpretacji innych parametrów za wyjątkiem interpretacji parametru e^{β_0}, gdzie β_0 jest wyrazem wolnym. Wartość e^{β_0} jest to szansa wystąpienia zjawiska w grupie referencyjnej (wówczas wszystkie zmienne i iloczyny tych zmiennych są równe zero).

Interpretacja parametrów modelu jest odmienna dla pojedynczej zmiennej, iloczynu dwóch zmiennych, iloczynu trzech zmiennych itd., co zostanie przedstawione na przykładzie modelu (11) z trzema zmiennymi uwzględniającego wszystkie efekty interakcji. W modelu tym dla zmiennej nie będącej iloczynem np. X_1 parametr e^{β_1} można otrzymać wyznaczając iloraz szans dla jednostkowego przyrostu zmiennej X_1:

$$\psi(x_1, x_2, x_3;\ 1,\ 0,\ 0) = \frac{\gamma(x_1 + 1, x_2, x_3)}{\gamma(x_1, x_2, x_3)} = \exp(\beta_1 + \beta_4 x_2 + \beta_5 x_3 + \beta_7 x_2 x_3) \quad (12)$$

pod warunkiem, że $X_2 = X_3 = 0$ (zmienne będące moderatorami zmiennej X_1 są równe zero).

Natomiast dla iloczynu dwóch zmiennych np. X_1X_2 z modelu (11) parametr e^{β_4} można uzyskać dzieląc ilorazy szans. Jeżeli iloraz szans (12), w którym w miejsce wartości x_2 wstawiono $x_2 + 1$ zostanie podzielony przez iloraz szans (12) z drugą zmienną na poziomie x_2, to otrzymany zostanie następujący iloraz ilorazów szans:

$$\begin{aligned}&\frac{\psi(x_1, x_2 + 1,\ x_3;\ 1,\ 0,\ 0)}{\psi(x_1, x_2,\ x_3;\ 1,\ 0,\ 0)} = \\ &= \frac{\exp\left(\beta_1 + \beta_4(x_2 + 1) + \beta_5 x_3 + \beta_7(x_2 + 1)x_3\right)}{\exp\left(\beta_1 + \beta_4 x_2 + \beta_5 x_3 + \beta_7 x_2 x_3\right)} = \exp\left(\beta_4 + \beta_7 x_3\right),\end{aligned} \tag{13}$$

który przyjmuje wartość e^{β_4} dla $X_3 = 0$. Można wykazać, że $\frac{\psi(x_1, x_2 + 1,\ x_3;\ 1,\ 0,\ 0)}{\psi(x_1, x_2,\ x_3;\ 1,\ 0,\ 0)} = \frac{\psi(x_1 + 1, x_2,\ x_3;\ 0,\ 1,\ 0)}{\psi(x_1, x_2,\ x_3;\ 0,\ 1,\ 0)}$, czyli przy wyznaczaniu ilorazu ilorazów szans (13) nie ma znaczenia w jakiej kolejności badane są jednostkowe przyrosty zmiennej X_1 i X_2.

W przypadku iloczynu trzech zmiennych $X_1X_2X_3$ z modelu (11) parametr e^{β_7} można wyznaczyć jako iloraz ilorazu szans (13), w którym w miejsce wartości x_3 wstawiono $x_3 + 1$ podzielony przez iloraz ilorazu szans (13) z trzecią zmienną na poziomie x_3:

$$\frac{\dfrac{\psi(x_1, x_2 + 1,\ x_3 + 1;\ 1,\ 0,\ 0)}{\psi(x_1, x_2,\ x_3 + 1;\ 1,\ 0,\ 0)}}{\dfrac{\psi(x_1, x_2 + 1,\ x_3;\ 1,\ 0,\ 0)}{\psi(x_1, x_2,\ x_3;\ 1,\ 0,\ 0)}} = \frac{\exp\left(\beta_4 + \beta_7(x_3 + 1)\right)}{\exp\left(\beta_4 + \beta_7 x_3\right)} = \exp\left(\beta_7\right). \tag{14}$$

Podobnie, jak dla wyrażenia (13) można wykazać, że przy wyznaczaniu ilorazu ilorazów ilorazów szans (14) nie ma znaczenia w jakiej kolejności badane są jednostkowe przyrosty zmiennej X_1, X_2 i X_3.

Biorąc pod uwagę wyrażenia (12)-(14) można sformułować interpretacje parametrów e^{β} w modelach ze zmiennymi jakościowymi i ilościowymi. Interpretacja parametrów modelu w przypadku analizy interakcji między predyktorami jakościowymi reprezentowanymi przez zmienne zero-jedynkowe została przedstawiona poniżej w punktach 1)-3).

1) **Interpretacja parametru e^{β} dla zmiennej X nie będącej iloczynem zmiennych**

Parametr e^{β} jest równy ilorazowi szans dla grupy, w której $X = 1$ oraz grupy, w której $X = 0$, pod warunkiem, że zmienne będące moderatorami zmiennej X przyjmują wartość zero, czyli zmienne występujące w iloczynach z analizowaną zmienną są równe zero (przy pozostałych zmiennych jednakowych w obu grupach).

2) **Interpretacja parametru e^{β} dla zmiennej *XZ* będącej iloczynem dwóch zmiennych**

Obecność interakcji stopnia drugiego odzwierciedla się w zróżnicowaniu ilorazów szans. Porównania ilorazów szans można dokonać wyznaczając ich iloraz i sprawdzając, czy różni się statystycznie istotnie od 1. Jeżeli zmienna Z zostanie potraktowana jako moderator zmiennej X, to współczynnik e^{β} jest równy ilorazowi dwóch ilorazów szans otrzymanemu poprzez podzielenie ilorazu szans dla grupy, w której $X = 1$ oraz grupy, w której $X = 0$ pod warunkiem, że $Z = 1$ przez iloraz tych szans pod warunkiem, że $Z = 0$ (przy pozostałych moderatorach zmiennej X równych zero oraz ustalonej wartości pozostałych zmiennych niezależnych).

3) **Interpretacja parametru e^{β} dla zmiennej *QXZ* będącej iloczynem trzech zmiennych**

Tak, jak obecność interakcji stopnia drugiego można stwierdzić dzieląc ilorazy szans, tak obecność interakcji stopnia trzeciego można zaobserwować dzieląc ilorazy ilorazów szans. Niech zmienne Q oraz Z będą moderatorami zmiennej X, to wartość e^{β} można uzyskać dzieląc iloraz ilorazu szans otrzymany tak, jak przy badaniu interakcji stopnia drugiego między zmienną X oraz Z, lecz pod warunkiem, że $Q = 1$ przez iloraz ilorazu szans otrzymany analogicznie, tym razem pod warunkiem, że $Q = 0$ (przy pozostałych moderatorach zmiennej X równych zero oraz ustalonej wartości pozostałych zmiennych niezależnych).

W sytuacji, gdy analizowana jest interakcja między predyktorami jakościowymi i ilościowymi lub interakcja między predyktorami tylko ilościowymi interpretacje 1)-3) należy nieco zmienić. W przypadku zmiennej ilościowej w powyższych interpretacjach zastępuje się grupy zakodowane jako 0 oraz jako 1 przez dwie grupy, dla których zmienna ilościowa różni się o 1 jednostkę. (por. (12)-(14) i [6]).

4. ANALIZA RYZYKA ZGONU Z WYKORZYSTANIEM MODELU LOGITOWEGO

Analiza dotyczy ryzyka zgonu osób starszych (w wieku od 60 lat) w województwie pomorskim w roku 2009. Model zbudowano na podstawie następujących danych pochodzących z GUS:

1. liczba osób zmarłych w roku 2009,
2. liczba osób żyjących na końcu roku 2008 oraz 2009

sklasyfikowanych jednocześnie według roku urodzenia, wieku ukończonego, płci, miejsca zamieszkania (miasto/wieś). Wiek osób przyjęto na moment 31 XII 2008, czyli w roku 2009 obserwowane były generacje osób urodzonych w roku 1948 i wcześniej, przy czym uwzględniono poprawkę na liczbę osób migrujących (por. [12, s. 36-37]). W okresie roku kalendarzowego obserwacji poddano 379 tys. mieszkańców województwa pomorskiego. Jednoczesna klasyfikacja osób według wszystkich cech badania po-

zwoliła na uzyskanie indywidualnych danych dla każdej jednostki.[5] Parametry modelu wyznaczono metodą największej wiarygodności przy użyciu modułu „Uogólnione modele liniowe" zawartego w pakiecie komputerowym *Statistica 8.0*.

Zróżnicowanie ryzyka zgonu w zależności od płci, wieku, miejsca zamieszkania można wstępnie zaobserwować porównując empiryczne (nie poddane modelowaniu) wartości prawdopodobieństwa oraz szansy zgonu w grupach osób sklasyfikowanych pod względem tych cech (patrz tabela 1).

Tabela 1

Empiryczne wartości prawdopodobieństwa oraz szansy zgonu według płci, miejsca zamieszkania i 5-letnich grup wieku w województwie pomorskim w 2009 roku

Wiek	Prawdopodobieństwo zgonu				Szansa zgonu			
	Miasta		Wieś		Miasta		Wieś	
	Mężczyźni	Kobiety	Mężczyźni	Kobiety	Mężczyźni	Kobiety	Mężczyźni	Kobiety
60-64	0,01957	0,00890	0,02257	0,01075	0,01996	0,00898	0,02309	0,01087
65-69	0,03073	0,01393	0,03274	0,01605	0,03170	0,01413	0,03385	0,01632
70-74	0,03997	0,02106	0,05221	0,02248	0,04164	0,02151	0,05508	0,02300
75-79	0,06355	0,03681	0,07222	0,04058	0,06786	0,03821	0,07784	0,04230
80-84	0,09416	0,06926	0,10776	0,07484	0,10394	0,07442	0,12078	0,08089
85+	0,16845	0,13944	0,18120	0,15395	0,20257	0,16203	0,22130	0,18197

Źródło: obliczenia własne na podstawie danych GUS.

Regresja logistyczna umożliwiła znalezienie statystycznie istotnych predyktorów ryzyka zgonu oraz odkrycie efektów interakcji między tymi predyktorami. Zmienna objaśniana przyjęła postać

$$Y = \begin{cases} 1 & \text{zgon} \\ 0 & \text{przezycie} \end{cases} . \tag{15}$$

Dwa predyktory jakościowe płeć i miejsce zamieszkania zostały wprowadzone do modelu przy pomocy zmiennych zero-jedynkowych. Grupę referencyjną utworzyły kobiety mieszkające w mieście. Zmienną wiek w analizie uwzględniono na dwa sposoby:

[5] Przy użyciu metody największej wiarygodności model estymowany na podstawie danych indywidualnych jest identyczny z modelem estymowanym na podstawie danych pogrupowanych jednocześnie według wszystkich wariantów cech występujących dla danych indywidualnych [7, s. 105-106].

1. jako zmienną skategoryzowaną – wyodrębniono sześć kategorii wieku (patrz tabela 1)[6] reprezentowanych w modelu przez pięć zmiennych zero-jedynkowych[7] (najmłodsza grupa wieku od 60 do 64 lat została grupą referencyjną),
2. jako zmienną ciągłą minus 60 (w tym przypadku 60-latkowie utworzyli grupę referencyjną).

Powyższe dwa podejścia mają swoje ograniczenia. Wprowadzenie predyktora wiek do modelu jako zmiennej ilościowej wymaga założenia, aby logit był linową funkcją wieku, a tym samym oznacza to, że każdy wzrost wieku o rok powoduje taki sam procentowy wzrost szansy zgonu (stały iloraz szans dla jednakowych przyrostów wieku). Założenie to można ominąć kategoryzując zmienną i wprowadzając ją do modelu jako zespół zmiennych zero-jedynkowych. Dla każdej zmiennej zero-jedynkowej oszacowany zostaje odrębny współczynnik regresji, więc ilorazy szans w sąsiednich grupach wieku nie muszą być stałe. Jednakże wyodrębnienie grup wieku powoduje utratę szczegółowości danych.

4.1 WYNIKI ESTYMACJI MODELU LOGITOWEGO BEZ UWZGLĘDNIENIA EFEKTÓW INTERAKCJI

Konstrukcja dwumianowego modelu logitowego pozwoliła na identyfikację predyktorów ryzyka zgonu. Wszystkie analizowane zmienne objaśniające okazały się statystycznie istotne zarówno w wariancie z wiekiem jako zmienną skategoryzowaną (tabela 2) oraz w wariancie z wiekiem jako zmienną ciągłą (tabela 3). Występowanie zróżnicowania ryzyka zgonu w zależności od płci, wieku, miejsca zamieszkania można stwierdzić sprawdzając, czy współczynnik regresji β różni się statystycznie istotnie od 0 na poziomie istotności $\alpha = 0,05$, co jest równoznaczne z tym, że iloraz szans równy w modelu logitowym e^{β} różni się statystycznie istotnie od 1, a także tym, że 95% przedział ufności dla e^{β} nie zawiera 1 (por. np. [5]).

W przypadku wyrazu wolnego współczynnik e^{β} równa się szansie wystąpienia zgonu w grupie referencyjnej, tzn. grupie kobiet w wieku 60-64 lata (tabela 2) lub w wieku 60 lat (tabela 3) zamieszkujących w miastach (por. wartości z tabeli 2 i 3 z rzeczywistą wartością z tabeli 1). Jest to grupa najniższego ryzyka w badanej populacji (potwierdzają to dodatnie współczynniki regresji β przy zmiennych objaśniających). Analizując ilorazy szans e^{β} z tablicy 2 i 3 można stwierdzić, że oba modele wskazują, że szansa zgonu mężczyzn jest większa o około 71%-73% niż szansa zgonu kobiet, natomiast szansa zgonu mieszkańca wsi jest większa o około 14% niż szansa zgonu mieszkańca miasta. Dla zmiennej skategoryzowanej wiek oszacowano ilorazy szans zgonu e^{β} w danej grupie wieku w stosunku do grupy referencyjnej 60-64 lata, np.

[6] Są to grupy wieku stosowane tradycyjnie do prezentacji danych demograficznych. Problem wyodrębnienia grup wieku, które są zróżnicowane ze względu na prawdopodobieństwo zgonu jest złożonym tematem (dyskusji można poddać wybór liczby grup, czy granic przedziałów wieku, tzw. punktów odcięcia).

[7] Inny sposób kodowania grup wieku zastosowano m. in. w [9, s. 148-151].

Tabela 2

Wyniki estymacji modelu logitowego z zero-jedynkowymi zmiennymi objaśniającymi bez uwzględnienia efektów interakcji

	Ocena parametru β	Błąd standardowy	e^{β}	95% przedział dla e^{β}	
Wyraz wolny	-4,5421	0,0281	0,0107	0,0101	0,0113
Płeć (1-mężczyzna, 0-kobieta)	0,5381	0,0174	1,7127	1,6552	1,7722
Miejsce (1-wieś, 0-miasto)	0,1324	0,0191	1,1415	1,0997	1,1850
Wiek 65-69	0,4341	0,0354	1,5435	1,4401	1,6545
Wiek 70-74	0,7888	0,0334	2,2007	2,0612	2,3496
Wiek 75-79	1,2922	0,0321	3,6409	3,4190	3,8771
Wiek 80-84	1,8565	0,0322	6,4014	6,0098	6,8185
Wiek 85+	2,6186	0,0320	13,7166	12,8833	14,6037
test ilorazu wiarygodności: *p-value* = 0,0000 pseudo-R^2 Vealla-Zimmermanna = 0,1040			czułość modelu = 58,9% swoistość modelu = 74,7%		

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

Tabela 3

Wyniki estymacji modelu logitowego z zero-jedynkowymi zmiennymi płeć i miejsce zamieszkania oraz zmienną ciągłą wiek bez uwzględnienia efektów interakcji

	Ocena parametru β	Błąd standardowy	e^{β}	95% przedział dla e^{β}	
Wyraz wolny	-4,8650	0,0227	0,0077	0,0074	0,0081
Płeć (1-mężczyzna, 0-kobieta)	0,5467	0,0175	1,7275	1,6694	1,7877
Miejsce (1-wieś, 0-miasto)	0,1358	0,0191	1,1454	1,1033	1,1891
Wiek	0,1001	0,0010	1,1053	1,1031	1,1075
test ilorazu wiarygodności: *p-value* = 0,0000 pseudo-R^2 Vealla-Zimmermanna = 0,1070			czułość modelu = 66,3% swoistość modelu = 67,8%		

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

szansa zgonu w grupie 65-69 lat jest większa o 54% niż w grupie 60-64 lata, szansa zgonu w grupie 70-74 lat jest większa o 120% niż w grupie 60-64 lata itd. Na tej podstawie można także obliczyć, iż w stosunku do sąsiedniej młodszej grupy szansa zgonu jest większa w grupie: 70-74 lat o 43%, 75-79 lat o 65%, 80-84 lat o 76%, 85+ lat o 114%. Natomiast w modelu przedstawionym w tabeli 3 zwiększenie wieku o rok powoduje wzrost szansy zgonu o 10,5%, czyli zwiększenie wieku o 5 lat powoduje wzrost szansy zgonu o 65% ($e^{5\beta} = 1,6496$).

4.2 WYNIKI ESTYMACJI MODELU Z UWZGLĘDNIENIEM EFEKTÓW INTERAKCJI

Płeć, miejsce zamieszkania i wiek są istotnymi predyktorami ryzyka zgonu, jednakże ich wpływ na zmienną objaśnianą może zależeć od wzajemnych interakcji. Uwzględnienie wszystkich interakcji (do stopnia 3 włącznie) w modelu ze wszystkimi zmiennymi zero-jedynkowymi powoduje, że prawdopodobieństwa zgonu oszacowane na podstawie takiego modelu są równe prawdopodobieństwom empirycznym dla danych pogrupowanych jednocześnie według płci, miejsca zamieszkania i grup wieku (wartości zmiennych objaśniających podzieliły w tym przypadku zbiorowość na $2 \times 2 \times 6 = 24$ grupy). Prowadzi to do rozbudowanego modelu (z 24 parametrami) mimo występowania tylko trzech predyktorów. Jednak nie wszystkie zmienne interakcyjne okazały się statystycznie istotne. W tabeli 4 przedstawiono model ze zmiennymi dychotomicznymi zawierający istotne statystycznie efekty interakcji. Budowa modelu metodą krokową

Tabela 4

Wyniki estymacji modelu logitowego z zero-jedynkowymi zmiennymi objaśniającymi z uwzględnieniem efektów interakcji

	Ocena parametru β	Błąd standardowy	e^{β}	95% przedział dla e^{β}	
Wyraz wolny	-4,6966	0,0437	0,0091	0,0084	0,0099
Płeć (1-mężczyzna, 0-kobieta)	0,7877	0,0538	2,1984	1,9783	2,4430
Miejsce (1-wieś, 0-miasto)	0,1287	0,0190	1,1374	1,0957	1,1806
Wiek 65-69	0,4411	0,0591	1,5545	1,3845	1,7453
Wiek 70-74	0,8407	0,0545	2,3181	2,0830	2,5796
Wiek 75-79	1,4243	0,0514	4,1550	3,7565	4,5957
Wiek 80-84	2,0859	0,0498	8,0519	7,3027	8,8780
Wiek 85+	2,8733	0,0486	17,6951	16,0863	19,4648
Płeć * Wiek 65-69	-0,0017	0,0738	0,9983	0,8638	1,1538
Płeć * Wiek 70-74	-0,0637	0,0691	0,9383	0,8194	1,0744
Płeć * Wiek 75-79	-0,2033	0,0662	0,8161	0,7168	0,9290
Płeć * Wiek 80-84	-0,4352	0,0668	0,6472	0,5677	0,7377
Płeć * Wiek 85+	-0,5717	0,0676	0,5646	0,4945	0,6446
test ilorazu wiarygodności: *p-value* = 0,0000 pseudo-R^2 Vealla-Zimmermanna = 0,1055			czułość modelu = 64,9% swoistość modelu = 68,9%		

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

postępującą oraz krokową wsteczną dała ten sam rezultat.[8] Statystycznie istotna okazała się interakcja płci z wiekiem, natomiast nieistotne okazały się interakcje stopnia drugiego miejsca zamieszkania z płcią oraz miejsca zamieszkania z wiekiem, jak też interakcja stopnia trzeciego między analizowanymi trzema predyktorami.[9]

Współczynnik e^{β} dla zmiennej płeć (tabela 4) mówi, że szansa zgonu mężczyzn jest około dwa razy większa niż szansa zgonu kobiet, ale tylko w wieku 60-64 lata (grupa referencyjna dla moderatora zmiennej płeć) zarówno na wsi, jak i w mieście (miejsce zamieszkania nie jest moderatorem zmiennej płeć). Interpretacja współczynnika e^{β} dla zmiennej miejsce zamieszkania nie zmieniła się, gdyż zmienna ta nie posiada moderatora (szansa zgonu mieszkańca wsi jest większa o około 14% niż szansa zgonu mieszkańca miasta). W przypadku zmiennych zero-jedynkowych reprezentujących grupy wieku współczynnik e^{β} jest równy ilorazowi szans zgonu w danej grupie wieku oraz grupie referencyjnej w wieku 60-64 lata, ale tylko dla kobiet (moderator zmiennej wiek równa się wówczas zero), np. szansa zgonu kobiet w wieku 65-69 lat jest o 55% większa niż kobiet w wieku 60-64 lata (zarówno na wsi, jak i w mieście). Współczynnik e^{β} przy iloczynie zmiennej płeć oraz zmiennej zero-jedynkowej reprezentującej grupę wieku interpretowany jest jako iloraz dwóch ilorazów szans, np. iloraz szans zgonu mężczyzn i kobiet w wieku 85 lat i więcej jest mniejszy o 43,5% od ilorazu szans mężczyzn i kobiet w wieku 60-64 lata (iloraz szans zgonu mężczyzn i kobiet zmniejsza się z wiekiem, gdyż powyżej 60 lat maleje przewaga umieralności mężczyzn nad umieralnością kobiet).

W celu zobrazowania dopasowania modeli ze wszystkimi zmiennymi zero-jedynkowymi na wykresie 1 przedstawiono różnice względne (w %) między prawdopodobieństwem zgonu teoretycznym a prawdopodobieństwem empirycznym we wszystkich grupach wieku dla kobiet i mężczyzn mieszkających w miastach.[10] Model bez interakcji jest znacznie gorzej dopasowany: w zależności od wariantów analizowanych cech różnice względne wahały się od kilku do około 20%. W przypadku modelu ze wszystkimi interakcjami stopnia drugiego oraz modelu z interakcjami wyłącznie statystycznie istotnymi (tabela 4) różnice względne są znacznie mniejsze i nie przekraczają 5%, więc przyjęcie modelu zawierającego interakcje jedynie statystycznie istotne jest całkowicie zadawalające.[11]

---

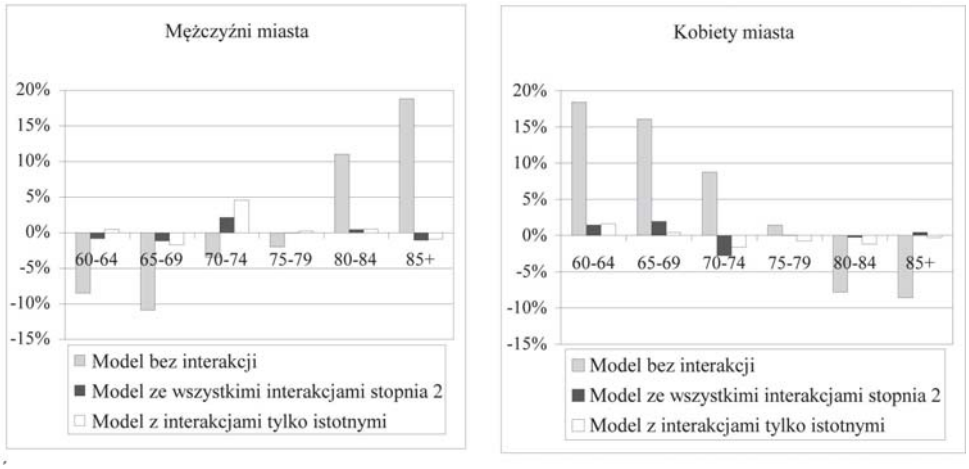
[8] Brak reguł przy przeszukiwaniu zbioru zmiennych może prowadzić do przypadkowych wyników, dlatego wykorzystuje się algorytmy do wyboru zmiennych objaśniających do modelu (*stepwise logistic regression*). (Por. [2] i [5])

[9] Zgodnie z kryterium informacyjnym Akaike'a oraz bayesowskim kryterium informacyjnym spośród modeli ze zmiennymi dychotomicznymi najlepszy okazał się również model opisany w tabeli 4.

[10] Dla osób mieszkających na wsi prawidłowości zaobserwowane w odchyleniach wartości teoretycznych od empirycznych były podobne jak dla mieszkańców miast (nieistotne interakcje miejsca zamieszkania z pozostałymi predyktorami).

[11] Na wykresie 1 nie zaprezentowano odchyleń dla modelu HWF ze wszystkimi efektami interakcji (w tym przypadku do stopnia trzeciego włącznie), gdyż dla takiego modelu odchylenia są równe zero (patrz uwaga w punkcie 3).

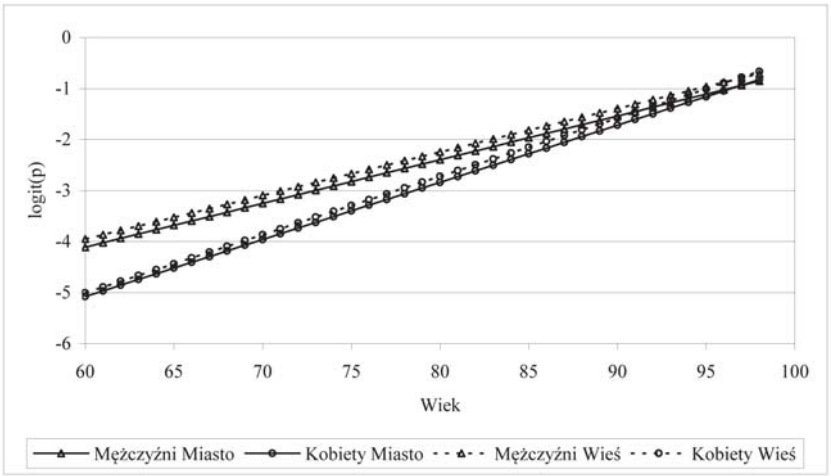
Wykres 1. Różnice względne w % między prawdopodobieństwem zgonu teoretycznym (oszacowanym na podstawie modelu z zero-jedynkowymi zmiennymi objaśniającymi) a prawdopodobieństwem empirycznym



Źródło: opracowanie własne.

W przypadku wprowadzenia do modelu predyktora wiek jako zmiennej ciągłej można graficznie przedstawić logit jako liniową funkcję wieku w grupach według płci i miejsca zamieszkania. Wykres 2 prezentuje przebieg tej funkcji oszacowanej na podstawie modelu zawierającego wszystkie efekty interakcji (czyli do stopnia trzeciego włącznie).

Wykres 2. Logit jako funkcja zmiennej ciągłej wiek w grupach według płci i miejsca zamieszkania – wyniki estymacji modelu zawierającego wszystkie efekty interakcji



Źródło: opracowanie własne.

Różne kąty nachylenia prostych w grupie kobiet i mężczyzn świadczą o występowaniu interakcji między predyktorami wiek i płeć. Natomiast zbliżony do równoległego przebieg prostych dla mężczyzn zamieszkałych w miastach i na wsi oraz równoległy przebieg prostych dla kobiet zamieszkałych w miastach i na wsi świadczy o nieistotnej sile interakcji między predyktorami wiek i miejsce zamieszkania (por. [4] i [6]). Zbliżone odległości między parami tych prostych świadczą także o niewielkiej sile interakcji między predyktorami płeć i miejsce zamieszkania. Procedura konstrukcji modelu zarówno krokowa postępująca oraz krokowa wsteczna potwierdziły powyższe obserwacje. Statystycznie istotne okazały się predyktory płeć, miejsce zamieszkania, wiek oraz iloczyn wieku i płci (tabela 5).[12]

Tabela 5

Wyniki estymacji modelu logitowego z zero-jedynkowymi zmiennymi płeć i miejsce zamieszkania oraz zmienną ciągłą wiek z uwzględnieniem efektów interakcji

	Ocena parametru β	Błąd standardowy	e^{β}	95% przedział dla e^{β}	
Wyraz wolny	-5,0962	0,0297	0,0061	0,0058	0,0065
Płeć (1-mężczyzna, 0-kobieta)	0,9927	0,0383	2,6984	2,5030	2,9090
Miejsce (1-wieś, 0-miasto)	0,1332	0,0191	1,1425	1,1005	1,1860
Wiek	0,1126	0,0014	1,1191	1,1160	1,1222
Wiek*Płeć	-0,0270	0,0021	0,9734	0,9695	0,9773
test ilorazu wiarygodności: *p-value* = 0,0000 pseudo-R^2 Vealla-Zimmermanna = 0,1088			czułość modelu = 68,1% swoistość modelu = 66,1%		

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

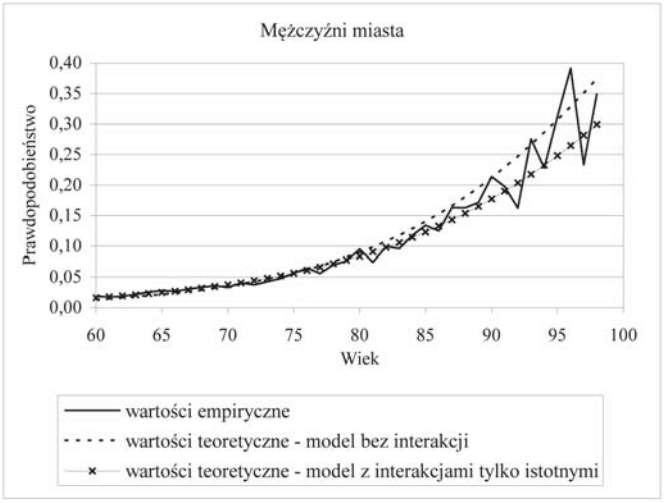
Współczynniki e^{β} dla zmiennej płeć (tabela 5) mówią, że szansa zgonu mężczyzn jest około 2,7 razy większa niż szansa zgonu kobiet, ale tylko w wieku 60 lat (grupa referencyjna dla moderatora zmiennej płeć) zarówno na wsi, jak i w mieście (miejsce zamieszkania nie jest moderatorem zmiennej płeć). Interpretacja współczynnika e^{β} dla zmiennej miejsce zamieszkania nie zmieniła się, gdyż zmienna ta nie posiada moderatora (szansa zgonu mieszkańca wsi jest większa o około 14% niż szansa zgonu mieszkańca miasta). W przypadku zmiennej wiek współczynnik e^{β} informuje, że wraz ze wzrostem wieku o jeden rok szansa zgonu rośnie o około 12%, ale tylko w grupie kobiet (moderator zmiennej wiek równa się wówczas zero). Współczynnik e^{β} przy iloczynie zmiennej płeć oraz wiek mówi, że jeżeli wiek wzrośnie o jeden rok, to iloraz szans zgonu mężczyzn i kobiet zmaleje o 2,7% (wraz z wiekiem maleje przewaga umieralności mężczyzn nad umieralnością kobiet).

Prawdopodobieństwa zgonu empiryczne i teoretyczne jako funkcje wieku dla mężczyzn mieszkających w miastach przedstawiono na wykresie 3, a dla kobiet mieszkają-

[12] Kryterium informacyjne Akaike'a oraz bayesowskie kryterium informacyjne również wskazały na model opisany w tabeli 5.

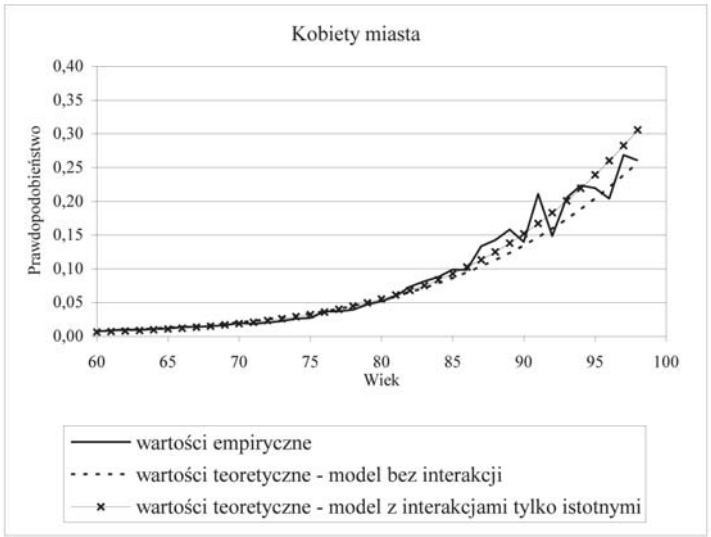
cych w miastach na wykresie 4.[13] Na wykresie dokonano porównania dla modelu bez interakcji oraz modelu z interakcjami tylko statystycznie istotnymi.

Wykres 3. Prawdopodobieństwa zgonu empiryczne i teoretyczne (oszacowane na podstawie modelu ze zmienną ciągłą wiek) w przypadku mężczyzn zamieszkujących w miastach



Źródło: opracowanie własne.

Wykres 4. Prawdopodobieństwa zgonu empiryczne i teoretyczne (oszacowane na podstawie modelu ze zmienną ciągłą wiek) w przypadku kobiet zamieszkujących w miastach



Źródło: opracowanie własne.

[13] Analogiczne prawidłowości można zaobserwować dla mieszkańców wsi (nieistotne interakcje miejsca zamieszkania z pozostałymi predyktorami).

Empiryczne prawdopodobieństwa zgonu traktowane jako funkcja wieku charakteryzują się nieregularnym przebiegiem, w szczególności dla najstarszych roczników (mniejsza liczba danych). Do ich wygładzenia (wyrównania) można wykorzystać modele logitowe. W przypadku modelu bez interakcji znaki odchyleń wartości empirycznych od teoretycznych tworzą zbyt długie serie. W przypadku mężczyzn model bez interakcji przeszacowuje prawdopodobieństwa zgonu (długie serie punktów empirycznych pod wykresem oszacowanej funkcji (wykres 3)), natomiast w przypadku kobiet model bez interakcji niedoszacowuje prawdopodobieństwa zgonu (długie serie punktów empirycznych nad wykresem oszacowanej funkcji (wykres 4))[14]. Uwzględnienie interakcji powoduje, że odchylenia wartości empirycznych od teoretycznych można uznać za losowe[15].

Ponieważ część efektów interakcji okazała się statystycznie nieistotna, więc przebieg funkcji ze wszystkimi zmiennymi interakcyjnymi oraz z wybranymi zmiennymi interakcyjnymi jest zbliżony. W praktyce poszukuje się funkcji jak najlepiej dopasowanej i jednocześnie o możliwie najmniejszej liczbie parametrów, więc jako model analityczny (parametryczny) ludzkiego procesu wymieralności wystarczy przyjąć model jedynie z interakcjami statystycznie istotnymi.

5. PODSUMOWANIE

W niniejszym artykule przedstawiono propozycję zastosowania modelu logitowego w analizie zróżnicowania ryzyka zgonu. Regresja logistyczna pozwoliła na oszacowanie prawdopodobieństwa zgonu w zależności od poziomu zmiennych objaśniających: płeć, miejsce zamieszkania oraz wiek. Zaletą modelu logitowego jest to, że obok ilościowych zmiennych niezależnych, mogą występować w modelu także jakościowe zmienne niezależne zakodowane w odpowiedni sposób.

W zależności od przeznaczenia modelu zmienną ilościową można skategoryzować, a następnie otrzymane kategorie zakodować np. za pomocą zmiennych zerojedynkowych. Model z wiekiem jako zmienną skategoryzowaną (tabela 2 i 4) pozwala na oszacowanie odrębnych współczynników regresji dla grup wieku, co pozwala uniknąć założenia o stałym ilorazie szans dla jednakowych przyrostów wieku i może być cenne dla oceny ryzyka np. metodą scoringową. Natomiast model z wiekiem jako zmienną ciągłą (tabela 3 i 5) może znaleźć zastosowanie do wygładzania ciągu empirycznych prawdopodobieństw zgonu (patrz wykres 3 i 4).

Wprowadzenie zmiennych interakcyjnych w postaci iloczynu zmiennych pozwoliło na wykrycie efektów interakcji między zmiennymi objaśniającymi, przy czym statystycznie istotna okazała się interakcja płci z wiekiem, natomiast nieistotne okazały

[14] W modelu bez interakcji uwzględniony jest stały (uśredniony) poziom nadumieralności mężczyzn w stosunku do kobiet, lecz nadumieralność mężczyzn maleje po 60 roku życia.

[15] Innym sposobem rozwiązania tego problemu byłoby oszacowanie modeli odrębnych dla mężczyzn i kobiet. Jednakże budowa modelu logitowego ze zmiennymi interakcyjnymi pozwala dodatkowo na ocenę siły i kierunku efektu interakcji.

się interakcje stopnia drugiego miejsca zamieszkania z płcią oraz miejsca zamieszkania z wiekiem, jak też interakcja stopnia trzeciego między analizowanymi trzema predyktorami.

Szczególną uwagę w artykule zwrócono na możliwość interpretacji parametrów e^{β}. Interpretacja ta zależy od sposobu kodowania zmiennych oraz od tego, czy współczynnik regresji β występuje przy zmiennej nie będącej iloczynem, czy też przy zmiennej będącej iloczynem, jak też od tego ile zmiennych występuje w iloczynie. Do interpretacji parametru e^{β} oszacowanego dla zmiennej nie będącej iloczynem wykorzystuje się pojęcie ilorazu szans, dla zmiennej będącej iloczynem dwóch zmiennych wykorzystuje się pojęcie ilorazu ilorazu szans, dla zmiennej będącej iloczynem trzech zmiennych wykorzystuje się pojęcie ilorazu ilorazu ilorazu szans itd. W artykule rozważone zostały przypadki interakcji między predyktorami jakościowymi, między predyktorem jakościowym i ilościowym oraz między predyktorami ilościowymi.

Model logitowy pozwala na uwzględnienie cech demograficzno-społecznych oraz efektu interakcji między tymi cechami przy szacowaniu prawdopodobieństwa zgonu, co może znaleźć praktyczne zastosowania, np. w dziedzinie ubezpieczeń, planach emerytalnych, czy medycynie.

Uniwersytet Gdański

LITERATURA

[1] Agresti A., [2002], *Categorical Data Analysis*, John Wiley & Sons, New Jersey.
[2] Christensen R. Ch., [1997], *Log-linear models and logistic regression*, Springer, New York.
[3] Gruszczyński M., [2010], *Modele zmiennych jakościowych dwumianowych*, W: *Mikroekonometria. Modele i metody analizy danych indywidualnych,* pod red. Gruszczyński M., Wolters Kluwer Polska, Warszawa.
[4] Harrell F., [2001], *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer-Verlag, New York.
[5] Hosmer D., Lemeshow S., [2000], *Applied Logistic Regression*, John Wiley & Sons, New Jersey.
[6] Jaccard J., [2001], *Interaction Effects in Logistic Regression*, Sage University Papers, Series: „Quantitative Applications in the Social Sciences", 07-135, Thousand Oaks.
[7] Jong de P., Heller G. Z., [2008], *Generalized Linear Models for Insurance Data*, Cambridge University Press, Cambridge.
[8] Maddala G. S., [1983], *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press, Cambridge.
[9] Mazurek E., [2000], *Model logitowy*, W: *Metody oceny i porządkowania ryzyka w ubezpieczeniach życiowych*, pod red. Ostasiewicz S., Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.
[10] McCullagh P., Nelder J. A., [1989], *Generalized Linear Models*, Chapman & Hall, London.
[11] Tabeau E., Willekens F., van Poppel F., [2002], *Parameterization as a tool in analyzing age, period and cohort effects on mortality: A case study of the Netherlands*, W: *The Life Table. Modelling Survival and Death*, pod red. Wunsch G., Mouchart M., Duchene J., Kluwer Academic Publishers, Dordrecht.
[12] *Trwanie życia w 2009 r*., [2010], Główny Urząd Statystyczny, Warszawa.

EFEKTY INTERAKCJI MIĘDZY ZMIENNYMI OBJAŚNIAJĄCYMI W MODELU LOGITOWYM W ANALIZIE ZRÓŻNICOWANIA RYZYKA ZGONU

S t r e s z c z e n i e

Celem pracy jest identyfikacja predyktorów ryzyka zgonu oraz zbadanie efektów interakcji pomiędzy nimi. W artykule wykorzystano model regresji logistycznej do oszacowania prawdopodobieństw zgonu osób starszych (w wieku od 60 lat) w województwie pomorskim w 2009 roku. Jako predyktory ryzyka zgonu przyjęto: wiek, płeć oraz miejsce zamieszkania (miasto/wieś). W modelu uwzględniono wiek na dwa sposoby: jako zmienną ciągłą oraz jako zmienną skategoryzowaną. Przeprowadzono analizę efektów interakcji między predyktorami poprzez wprowadzenie do modelu iloczynu zmiennych objaśniających. Szczególną uwagę zwrócono na interpretację współczynników w modelu zawierającym zmienne interakcyjne. Rozważono przypadki interakcji między predyktorami jakościowymi, między predyktorem jakościowym i ilościowym oraz między predyktorami ilościowymi. Statystycznie istotna okazała się interakcja płci z wiekiem.

Słowa kluczowe: regresja logistyczna, efekt interakcji, prawdopodobieństwo zgonu

INTERACTION EFFECTS BETWEEN PREDICTOR VARIABLES IN A LOGISTIC MODEL IN AN ANALYSIS OF THE DIVERSITY OF DEATH RISK

S u m m a r y

The aims of this paper include the identification of predictors of death risk and the examination of interaction effects between them. In this study, a logistic regression model is used to estimate death probability at old age (above 60) in the Pomorskie Voivodship in 2009. The following risk factors of death are considered: age, gender and place of residence (urban/rural areas). In the model, age is treated both as a continuous variable and as a categorical variable. The paper presents an analysis of interaction effects between predictors with the use of product terms in the logistic regression model. The emphasis is on the interpretation of the coefficients of the interactive logistic model. The study includes cases of interactions between qualitative predictors, between qualitative and quantitative predictors, and between quantitative predictors. It appears that the interaction between gender and age is statistically significant.

Key words: logistic regression, interaction effect, death probability