PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 — 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Cyprian Kozyra
Akademia Ekonomiczna we Wrocławiu

ZASTOSOWANIE REGRESJI LOGISTYCZNEJ DO ANALIZY DANYCH WIELOMIANOWYCH DOTYCZĄCYCH SAMOOCENY ZDROWIA MIESZKAŃCÓW TORUNIA

1. Wstęp

Pytanie o ocenę własnego zdrowia jest stawiane w ramach badań zdrowotności społeczeństw, zalecanych przez Światową Organizację Zdrowia WHO (por. [*Stan zdrowia...* 1997; *Stan zdrowia...* 2001]). Ogólnopolskie badanie stanu zdrowia ludności przeprowadzono w 1996 r., a jego wyniki opublikowano w opracowaniach GUS (m.in. [*Stan zdrowia...* 1997]) oraz w Internecie (zob. http://www.stat.gov.pl/serwis/nieregularne/zdrowie/index.htm). Ankietowani odpowiadali w tym badaniu na pytanie o postrzeganie własnego zdrowia, wybierając jedną z pięciu ocen: „bardzo dobre", „dobre", „takie sobie (ani dobre, ani złe)", „złe", „bardzo złe". W badaniu wzięła udział reprezentatywna próba losowa ludności Polski, licząca przeszło 62 tys. osób. Opublikowane wyniki są jednak uogólnione na całą populację Polski za pomocą specjalnych mnożników wagowych. Uniemożliwia to dokonanie analizy istotności oszacowanych modeli regresji logistycznej. Z tego względu w niniejszym artykule przeanalizowano dane pochodzące z innego badania, przeprowadzonego w Toruniu w 2000 r. w ramach programu CINDI Światowej Organizacji Zdrowia. W badaniu uczestniczyło 2048 respondentów. Raport z tego badania opublikowano w Internecie (zob. [*Stan zdrowia...* 2001]). Ankietowani mieszkańcy Torunia odpowiadali na pytanie o ocenę swojego stanu zdrowia, wybierając również jedną z pięciu odpowiedzi. Były one jednak inaczej sformułowane niż w badaniu dotyczącym całej Polski. Respondenci mieli do wyboru następujące odpowiedzi: „dobry", „dość dobry", „ani dobry, ani zły", „raczej zły", „zły". Odmienne sformułowanie kategorii odpowiedzi sprawia, że rozkłady odpowiedzi otrzymane w obu badanych populacjach (ludności Polski i mieszkańców Torunia) są nieporównywalne, gdyż np. najlepsza ocena

zdrowia w drugim badaniu jest sformułowana tak, jak druga kategoria w pierwszym badaniu. Dane z badania stanu zdrowia mieszkańców Torunia zostaną wykorzystane do modelowania zależności samooceny zdrowia Torunian od takich cech demograficznych, jak: wiek, dochód i wykształcenie.

2. Model regresji logistycznej dla porządkowych cech wielomianowych

Odpowiedzi na pytanie ankietowe o samoocenę zdrowia można traktować jako realizacje zmiennej losowej o rozkładzie wielomianowym, którego funkcja rozkładu prawdopodobieństwa wyraża się wzorem:

$$P(Y_1 = y_1,\ldots,Y_s = y_s) = \frac{n!}{y_1!\ldots y_s!} p_1^{y_1}\ldots p_s^{y_s}, \tag{1}$$

gdzie s oznacza liczbę możliwych kategorii odpowiedzi, Y_l oznacza zmienne losowe odpowiadające liczbie jednostek w kategorii l, y_l oznacza zaobserwowane liczebności dla poszczególnych kategorii, p_l oznacza prawdopodobieństwo, że losowa jednostka znajdzie się w kategorii l, a n oznacza ogólną liczbę jednostek. Opisane parametry rozkładu muszą spełniać warunki $\sum_{l=1}^{s} p_l = 1, \sum_{l=1}^{s} y_l = n$. Gdy liczba kategorii s jest równa 2, to wzór (1) określa rozkład dwumianowy. Kategorie odpowiedzi cechy modelowanej rozkładem wielomianowym mogą być ze swej natury uporządkowane (jak przy ocenie własnego zdrowia) lub nieuporządkowane, a w pierwszym przypadku cecha może kategoryzować ukrytą zmienną ciągłą lub też może być ściśle jawna (i zarazem ściśle kategorialna). Niekiedy wyróżnia się też cechy częściowo uporządkowane, np. gdy do uporządkowanych kategorii odpowiedzi doda się jako osobną kategorię nieudzielenie odpowiedzi. Dla cech o dwu kategoriach odpowiedzi nie ma jednak różnicy w postaci modelu regresji logistycznej (nazywanego też modelem analizy logitowej) dla różnego typu zmiennych kategorialnych.

Regresja logistyczna dla danych dwumianowych wyraża się wzorem:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_j. \tag{2}$$

Iloraz prawdopodobieństw $p/(1-p)$, który przyjmuje wartości od 0 do $+\infty$, jest nazywany szansą danego zdarzenia. Wzór (2) można przedstawić także jako :

$$F^{-1}(p) = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_j, \tag{3}$$

gdzie $F(x) = \dfrac{\exp(x)}{1+\exp(x)}$ jest dystrybuantą rozkładu logistycznego.

Niekiedy we wzorze (3) używane są dystrybuanty innych rozkładów, np. dystrybuanta rozkładu normalnego lub rozkładu wartości ekstremalnych Gumbela. Zastosowanie rozkładu logistycznego wynika przede wszystkim z teorii uogólnionego modelu liniowego, gdyż funkcja logit(p) jest kanoniczną funkcją łączącą dla rozkładu dwumianowego, który w kanonicznej postaci wykładniczej rodziny rozkładów można przedstawić jako:

$$P(Y=y)=\binom{n}{y}p^{y}(1-p)^{n-y}=\binom{n}{y}(1-p)^{n}\exp\left[y\ln\left(\frac{p}{1-p}\right)\right]. \tag{4}$$

Przyjmuje się zazwyczaj, że samoocena zdrowia jest cechą z uporządkowanymi kategoriami, która kategoryzuje ukrytą cechę ciągłą, jaką jest subiektywne odczucie stanu własnego zdrowia. Dlatego przedstawiony zostanie model regresji logistycznej, który ma zastosowanie dla cech tego typu. Zależność między ukrytą cechą ciągłą ξ a obserwowalną cechą kategorialną y, której wartościami są kolejne liczby naturalne, można przedstawić wzorem:

$$y=\begin{cases}1, & \text{gdy } \xi\le\tau_1,\\ \quad\vdots & \\ l, & \text{gdy } \tau_{l-1}<\xi\le\tau_l,\\ \quad\vdots & \\ s, & \text{gdy } \tau_{s-1}<\xi,\end{cases} \tag{5}$$

gdzie liczba kategorii s jest w tym przykładzie równa 5. Jeżeli przyjmie się, że ukryta cecha ξ zależy liniowo od pewnych zmiennych x_i: $\xi=\beta_0+\sum_{j=1}^{m}\beta_j X_j+\varepsilon$, wówczas dystrybuantę zmiennej Y można przedstawić wzorem (6):

$$\begin{aligned}F_Y(l)&=P(Y\le l)=P(\xi\le\tau_l)=P\left(\beta_0+\sum_{j=1}^{m}\beta_j X_j+\varepsilon\le\tau_l\right)=\\ &=P\left(\varepsilon\le\tau_l-\beta_0-\sum_{j=1}^{m}\beta_j X_j\right)=F_\varepsilon\left(\tau_l-\beta_0-\sum_{j=1}^{m}\beta_j X_j\right).\end{aligned} \tag{6}$$

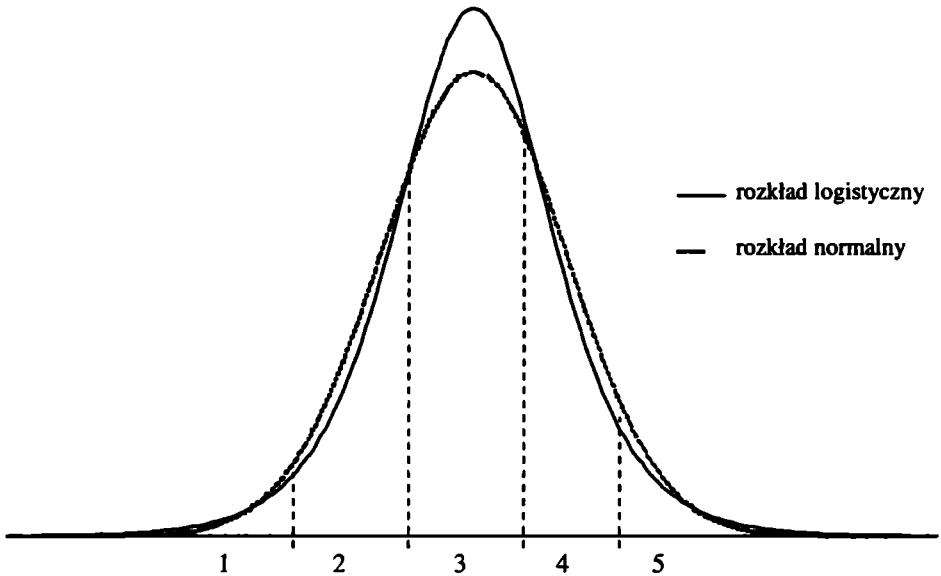
Rozkład zmiennej kategorialnej Y jest wyznaczany przez rozkład błędu losowego oraz pewne nieznane wartości progowe dla poszczególnych kategorii τ_l. Można to zilustrować schematycznie na rys. 1, na którym przedstawiono gęstość rozkładów normalnego i logistycznego o tej samej wariancji oraz wartości progowe τ_l. Rozkład logistyczny ma większą kurtozę i grubsze ogony niż rozkład normalny.

Oznaczając wartość dystrybuanty dla kategorii l przez g_l, otrzymuje się wzór (7):

$$F_Y(l)=\sum_{i=1}^{l}p_i=g_l=F_\varepsilon\left(\tau_l-\beta_0-\sum_{j=1}^{m}\beta_j x_j\right),\quad 1\le l\le s, \tag{7}$$

gdzie p_i oznacza prawdopodobieństwo znalezienia się w kategorii i. Po przyjęciu rozkładu logistycznego błędu losowego i oznaczeniu $\alpha_l = \beta_0 - \tau_l$ otrzymuje się równanie regresji (8) nazywane kumulacyjnym modelem regresji logistycznej:

$$\text{logit}(1-g_l) = \ln\left(\frac{1-g_l}{g_l}\right) = \alpha_l + \sum_{j=1}^{m}\beta_j x_j, \quad l = 1,\ldots,s-1. \tag{8}$$



Rys. 1. Schematyczne przedstawienie kategoryzacji cechy ukrytej
Źródło: opracowanie własne.

Charakterystyczną cechą kumulacyjnego modelu logitowego stosowanego dla uporządkowanych danych kategorialnych jest to, że współczynniki regresji, stojące przy zmiennych objaśniających, są takie same w każdym z $s-1$ równań regresji (8) dla poszczególnych kategorii odpowiedzi l, co jest konsekwencją założenia o ukrytej zmiennej ciągłej kategoryzowanej przez zmienną jawną i przekształcenia (6). Hiperpłaszczyzny regresji wyznaczone przez kumulacyjny model logitowy są zatem równoległe. Model ten jest nazywany także modelem proporcjonalnych szans (ang. *proportional odds*), gdyż zmiana jednej zmiennej objaśniającej x_m o wartość t powoduje $\exp(\beta_m t)$-krotną zmianę szansy, że respondent znajdzie się w wyższej kategorii niż l, zgodnie ze wzorem (9):

$$\begin{aligned}\frac{1-g_l'}{g_l'} &= \exp\left(\alpha_l + \sum_{j=1}^{m-1}\beta_j x_j + \beta_m(x_m + t)\right) = \\ &= \exp\left(\alpha_l + \sum_{j=1}^{m}\beta_j x_j\right)\exp(\beta_m t) = \frac{1-g_l}{g_l}\exp(\beta_m t)\end{aligned}, \tag{9}$$

gdzie g_l' jest zmienionym prawdopodobieństwem, że respondent znajdzie się co najwyżej w kategorii l. Szanse dla poszczególnych kategorii zostają zmienione proporcjonalnie w stosunku do szans przed zmianą x_m o wartość t. Wartość $\exp(\beta_j)$ nazywana jest ilorazem szans (ang. *odds ratio*), gdyż jest ona równa stosunkowi zmienionej szansy do szansy przed zmianą przy zwiększeniu wartości zmiennej x_j o jednostkę.

Estymacja modelu regresji logistycznej jest przeprowadzana metodami numerycznymi przez znalezienie maksimum funkcji logarytmu wiarygodności. Funkcja logarytmu wiarygodności dla rozkładu wielomianowego z uporządkowanymi kategoriami ma postać:

$$\ln L(\alpha_1,\ldots,\alpha_s,\beta_1,\ldots\beta_m)=\sum_{i=1}^{k}\left(\ln\left(\frac{n_i!}{y_{i1}!\ldots y_{is}!}\right)+\sum_{l=1}^{s} y_{il}\ln(p_{il})\right)= \\ =\sum_{i=1}^{k}\left(\ln\left(\frac{n_i!}{y_{i1}!\ldots y_{is}!}\right)+\sum_{l=1}^{s} y_{il}\ln(g_{il}-g_{il-1})\right), \tag{10}$$

gdzie $g_{il}=1\Big/\left(1+\exp\left(\alpha_l+\sum_{j=1}^{m}\beta_j x_{ij}\right)\right)$ zgodnie ze wzorem (8). Estymację można przeprowadzić w pakietach statystycznych (np. w module „Visual Generalized Linear Model" programu *Statistica* lub wybierając polecenie Statistics/Regression/Ordinal Logistic Regression w programie *Minitab*) albo za pomocą dodatku *Solver* w arkuszu kalkulacyjnym *Excel*. Przy korzystaniu z pakietów statystycznych należy zwrócić uwagę, że estymują one równanie regresji logistycznej postaci $\text{logit}(g_l)=\alpha_l+\sum_{j=1}^{m}\beta_j x_j$, w którym oszacowane współczynniki będą równe co do modułu współczynnikom z równania (8), natomiast różnić się będą znakiem.

W celu oceny dopasowania modelu do danych empirycznych wykorzystuje się głównie statystykę ilorazu wiarygodności D, nazywaną też statystyką odchylenia (ang. *deviance*), daną wzorem (11):

$$D=2\sum_{i=1}^{k}\sum_{l=1}^{s} y_{il}\ln\left(\frac{y_{il}}{\hat{y}_{il}}\right), \tag{11}$$

gdzie $\hat{y}_{il}=n_i\hat{p}_{il}$ oznacza przewidywane liczebności oszacowane w modelu. Alternatywnie można stosować statystykę χ^2 testu zgodności Pearsona daną wzorem (12):

$$\chi^2=\sum_{i=1}^{k}\sum_{l=1}^{s}\frac{(y_{il}-\hat{y}_{il})^2}{\hat{y}_{il}}. \tag{12}$$

Obie statystyki mają dla kumulacyjnego modelu logitowego asymptotyczny rozkład chi-kwadrat o $(k-1)(s-1)-m$ stopniach swobody, jednak ich rozkład mo-

że się znacznie różnić od rozkładu chi-kwadrat w przypadku, gdy przewidywane liczebności są mniejsze od 5. Statystyka D jest stosowana częściej, gdyż umożliwia ona porównywanie dopasowania tzw. modeli zagnieżdżonych. Mówimy, że pewien model jest zagnieżdżony w innym modelu, jeśli występują w nich te same zmienne objaśniające, a część parametrów estymowanych w drugim modelu jest w pierwszym modelu ustalona (np. równa zeru) lub jest funkcją innych parametrów. Różnica statystyk D dla modelu o mniejszej liczbie estymowanych parametrów i modelu o większej ich liczbie ma asymptotyczny rozkład chi-kwadrat o liczbie stopni swobody równej różnicy liczb stopni swobody w obu modelach.

Założenie o równości wszystkich współczynników regresji jest sprawdzane przez oszacowanie modelu, w którym współczynniki regresji są estymowane dla wszystkich kategorii odpowiedzi. Można to wyrazić wzorem analogicznym do wzoru (8):

$$\operatorname{logit}(1-g_l) = \ln\left(\frac{1-g_l}{g_l}\right) = \alpha_l + \sum_{j=1}^{m}\beta_{lj}x_j, \quad l = 1,\ldots,s-1. \tag{13}$$

W rezultacie można otrzymać oddzielne parametry regresji logistycznej dla poszczególnych kategorii. Model ten jest zagnieżdżony w modelu proporcjonalnych szans i ma więcej estymowanych parametrów dla tych samych zmiennych objaśnianych. Statystyka D, dana wzorem (11), oraz statystyka χ^2, dana wzorem (12), mają asymptotyczny rozkład chi-kwadrat o $(k-1)(s-1)-m(s-1)$ stopniach swobody, różnica statystyk odchylenia dla modelu (8) i modelu (13) ma zaś rozkład chi-kwadrat o $m(s-1)-m$ stopniach swobody. Estymacja modelu (13) i weryfikacja założenia o równości współczynników regresji nie są jeszcze dostępne w standardowych procedurach większości pakietów statystycznych (*Statistica* lub *Minitab*), natomiast można ich dokonać np. za pomocą dodatku *Solver* arkusza kalkulacyjnego *Excel*.

3. Zastosowanie modelu regresji logistycznej do danych empirycznych

Przedstawiony kumulacyjny model logitowy zostanie zastosowany do analizy danych pochodzących z badania stanu zdrowia mieszkańców Torunia. W raporcie z tych badań zamieszczono trzy tabele kontyngencji przedstawiające zależność samooceny zdrowia respondentów od cech demograficznych: wieku, wykształcenia oraz dochodu (por. [*Stan zdrowia* ... 2001]). Należy zwrócić uwagę na znaczną liczbę brakujących danych – w każdej z tabel brakuje informacji na temat ok. jednej czwartej spośród 2048 respondentów. W raporcie nie podano jednak przyczyn i rozkładu tak znacznego braku danych, dlatego nie uwzględniono ich w analizie. Ze względu na niskie liczebności w ostatniej kategorii odpowiedzi we wszystkich przykładach połączono ostatnie dwie kategorie przy testowaniu dopasowania modeli.

Dane dotyczące zależności samooceny zdrowia są umieszczone w tab. 1. W tabeli 2 zamieszczono oszacowane parametry trzech modeli regresji logistycznej oraz statystyki dopasowania tych modeli. Pierwszy model – „zerowy" – charakteryzuje się tym, że

wszystkie współczynniki regresji β_i są równe zeru. Testowanie dopasowania tego modelu do danych odpowiada testowaniu hipotezy o niezależności obu zmiennych – wieku i samooceny zdrowia. Hipoteza ta zostaje odrzucona, a wprowadzenie zależności liniowej, określonej w modelu proporcjonalnych szans (8), istotnie poprawia dopasowanie. Jednak model ten ciągle zbyt słabo pasuje do danych. Przy porównaniu wartości współczynników regresji tego modelu z wartościami modelu (13) – „pełnego" – można spostrzec, że mimo zwiększania się wieku respondentów, są oni mniej skłonni oceniać własne zdrowie jako gorsze od „dobrego" niż jako gorsze od kolejnych kategorii odpowiedzi. Ogólnie jednak można stwierdzić, że ocena własnego zdrowia pogarsza się wraz z wiekiem ankietowanych (co nie jest zaskakującym wnioskiem).

Tabela 1. Stan zdrowia w zależności od wieku

Grupy wieku	Stan zdrowia					Razem
	dobry	dość dobry	ani dobry, ani zły	raczej zły	zły	
	n	*n*	*n*	*n*	*n*	*n*
<25	192	83	14	5	1	295
25-34	130	98	30	5	5	268
35-44	174	119	41	10	1	345
45-54	102	114	76	36	5	333
55-64	69	90	78	38	9	284
Razem	667	504	239	94	21	1525

Źródło: [*Stan zdrowia...* 2001].

Tabela 2. Parametry i statystyki dopasowania modeli logitowych zależności samooceny zdrowia od wieku

Model zerowy								
Kategoria *l*	1	2	3	4		Statystyka *D*	Stat. Pearsona	*df*
α_l	0,2518	–1,1963	–2,5064	–4,2714		212,7542	208,3972	12
β_l	0	0	0	0	wartość *p*:	0,0000	0,0000	
Model proporcjonalnych szans								
Kategoria *l*	1	2	3	4		Statystyka *D*	Stat. Pearsona	*df*
α_l	–1,6583	–3,2496	–4,6411	–6,4368		29,2917	28,2289	11
β_l	0,0485	0,0485	0,0485	0,0485	wartość *p*:	0,0020	0,0030	
						Różnica stat. *D*		*df*
						183,4625		1
					wartość *p*:	0,0000		
Model pełny								
Kategoria *l*	1	2	3	4		Statystyka *D*	Stat. Pearsona	*df*
α_l	–1,4512	–3,7832	–5,3169	–6,1313		17,5483	17,1370	9
β_l	0,0428	0,0598	0,0625	0,0422	wartość *p*:	0,0408	0,0466	
						Różnica stat. *D*		*df*
						11,7435		2
					wartość *p*:	0,0028		

Źródło: opracowanie własne.

Dane dotyczące zależności samooceny zdrowia od dochodu respondentów zamieszczono w tab. 3. Ze względu na niskie liczebności połączono dwa ostatnie

przedziały klasowe dochodu zarówno przy estymacji, jak i przy ocenie dopasowania modelu. W tabe 4 przedstawiono wartości oszacowanych parametrów modeli i statystyk dopasowania. W tym przypadku wyższe dochody sprzyjają lepszym ocenom własnego zdrowia, a oprócz tego można stwierdzić liniową zależność oceny własnego zdrowia od dochodu w modelu proporcjonalnych szans, gdyż estymacja wszystkich współczynników regresji nie poprawia istotnie (na poziomie istotności 0,05) dopasowania modelu do danych. Można jednak zauważyć, że potrzebne jest znaczniejsze zmniejszenie dochodów, by respondenci oceniali swoje zdrowie jako gorsze od „dobrego", nie zaś jako gorsze od pozostałych kategorii.

Tabela 3. Stan zdrowia w zależności od przeciętnego dochodu netto

Dochód	Stan zdrowia					Razem
	dobry	dość dobry	ani dobry, ani zły	raczej zły	zły	
	n	*n*	*n*	*n*	*n*	*n*
<300	96	70	49	21	5	241
300-500	164	129	71	34	10	408
500-700	163	119	63	24	3	372
700-1000	83	92	34	7	1	217
1000-1500	73	32	9	6	0	120
>1500	38	26	8	0	2	74
Razem	617	468	234	92	21	1432

Źródło: [*Stan zdrowia*... 2001].

Tabela 4. Parametry i statystyki dopasowania modeli logitowych zależności samooceny zdrowia od dochodu

Model zerowy								
Kategoria *l*	1	2	3	4		Statystyka *D*	Stat. Pearsona	*df*
α_l	0,2783	–1,1400	–2,4572	–4,2075		45,7237	44,5145	12
β_l	0	0	0	0	wartość *p*:	0,0000	0,0000	
Model proporcjonalnych szans								
Kategoria *l*	1	2	3	4		Statystyka *D*	Stat. Pearsona	*df*
α_l	0,6695	–0,7668	–2,0960	–3,8513		20,9891	20,8043	11
β_l	–0,0610	–0,0610	–0,0610	–0,0610	wartość *p*:	0,0335	0,0355	
						Różnica stat. *D*		*df*
						24,7345		1
					wartość *p*:	0,0000		
Model pełny								
Kategoria *l*	1	2	3	4		Statystyka *D*	Stat. Pearsona	*df*
α_l	0,5857	–0,6030	–1,8524	–3,6202		15,7405	15,6758	9
β_l	–0,0487	–0,0910	–0,1074	–0,1055	wartość *p*:	0,0725	0,0740	
						Różnica stat. *D*		*df*
						5,2487		2
					wartość *p*:	0,0725		

Źródło: opracowanie własne.

Dane dotyczące zależności samooceny zdrowia od wykształcenia respondentów zamieszczono w tab. 5. Dwie ostatnie kategorie wykształcenia połączono przy estymacji i

ocenie dopasowania modelu ze względu na niskie zaobserwowane liczebności. Ponieważ wykształcenie jest kategorialną objaśniającą, jego kategorie zakodowano za pomocą trzech zmiennych zero-jedynkowych, zgodnie z tab. 6. Oszacowane współczynniki regresji umożliwiają porównanie samooceny zdrowia w odpowiedniej kategorii wykształcenia z bazową kategorią, którą jest wykształcenie podstawowe.

Tabela 5. Stan zdrowia w zależności od wykształcenia

Wykształcenie	Stan zdrowia					Razem
	dobry	dość dobry	ani dobry, ani zły	raczej zły	zły	
	n	*n*	*n*	*n*	*n*	*n*
Podstawowe	96	63	35	22	10	226
Zasadadnicze zawodowe	154	104	74	31	1	364
Średnie	275	194	94	33	8	604
Pomaturalne	29	25	11	2	0	67
Wyższe	106	111	24	6	2	249
Razem	660	497	238	94	21	1510

Źródło: [*Stan zdrowia* ... 2001].

Tabela 6. Kodowanie kategorii wykształcenia za pomocą zmiennych zero-jedynkowych

Kategoria wykształcenia	X_1	X_2	X_3
Podstawowe	0	0	0
Zasadnicze zawodowe	1	0	0
Średnie	0	1	0
Pomaturalne i wyższe	0	0	1

Źródło: opracowanie własne.

Tabela 7. Parametry i statystyki dopasowania modeli logitowych zależności samooceny zdrowia od wykształcenia

Model zerowy								
Kategoria *l*	1	2	3	4		Statystyka *D*	Stat. Pearsona	*df*
α_l	0,2530	–1,1871	–2,4957	–4,2613		45,0978	45,9326	9
β_{l1}	0	0	0	0	wartość *p*:	0,0000	0,0000	
β_{l2}	0	0	0	0				
β_{l3}	0	0	0	0				
Model proporcjonalnych szans								
Kategoria *l*	1	2	3	4		Statystyka *D*	Stat. Pearsona	*df*
α_l	0,4624	–0,9824	–2,2979	–4,0673		37,2282	37,5644	6
β_{l1}	–0,0657	–0,0657	–0,0657	–0,0657	wartość *p*:	0,0000	0,0000	
β_{l2}	–0,2791	–0,2791	–0,2791	–0,2791		Różnica stat. *D*		*df*
β_{l3}	–0,3578	–0,3578	–0,3578	–0,3578		7,8696		3
					wartość *p*:	0,0488		
Model pełny								
Kategoria *l*	1	2	3	4		Statystyka *D*	Stat. Pearsona	*df*
α_l	0,3032	–0,8642	–1,8021	–3,0727		0	0	0
β_{l1}	0,0070	–0,0253	–0,5373	–2,8217				
β_{l2}	–0,1239	–0,3811	–0,8176	–1,2381				
β_{l3}	–0,0100	–0,9312	–1,6189	–1,9835				

Źródło: opracowanie własne.

Wyniki przedstawione w tab. 7 wskazują na to, że model proporcjonalnych szans jest istotnie lepiej dopasowany do danych niż model niezależności samooceny zdrowia od wykształcenia. Model pełny jest w tym przykładzie modelem nasyconym (ang. *saturated*), gdyż jego statystyki dopasowania mają zero stopni swobody. Przy porównaniu wartości współczynników regresji tego modelu i modelu proporcjonalnych szans można spostrzec, że choć niższe wykształcenie ogólnie sprzyja gorszemu samopoczuciu zdrowotnemu, to na pewno nie występuje zależność liniowa. Różnice w wykształceniu nie wpływają tak znacznie na wybór przez respondentów gorszej kategorii i oceny zdrowia niż „dobre", jak na wybór gorszych ocen od pozostałych kategorii odpowiedzi. Spowodowane jest to zapewne zróżnicowanym wykształceniem młodzieży jeszcze się uczącej oraz ogólną poprawą wykształcenia społeczeństwa w ostatnich dziesięcioleciach, wskutek czego ludzie starsi o słabszym zdrowiu znajdują się w niższych kategoriach wykształcenia.

4. Zakończenie

Przedstawione modele logitowe są metodą analizy statystycznej w przypadku, gdy zmienna objaśniana jest kategorialna z uporządkowanymi kategoriami. Modele te są alternatywę wobec często stosowanych modeli regresji liniowej, w których kategoriom cechy naturalnej nadaje się wartości równe np. kolejnym liczbom naturalnym lub wynikające z porównania dystrybuanty empirycznej i dystrybuanty rozkładu zmiennej ukrytej. Podejście przedstawione w artykule jest bardziej zaawansowane teoretycznie i stanowi rozszerzenie dobrze znanych w polskiej literaturze modeli regresji logistycznej dla danych dwumianowych (por. [Jajuga 1990; Mazurek 1999]). Szersze omówienie modeli logistycznych dla wielomianowych cech kategorialnych można znaleźć w literaturze zagranicznej (por. [Agresti 1990; Cramer 2003; Rodriguez]). Przedstawione w artykule podejście może być stosowane także w dziedzinach innych niż badania zdrowotności, np. w badaniach marketingowych. Pewną słabością przeanalizowanych przykładów jest modelowanie zależności samooceny zdrowia tylko od jednej zmiennej objaśniającej. W Internecie nie są dostępne dane umożliwiające badanie wpływu większej liczby cech w jednym modelu. Niemniej można wnioski wyciągnięte z dostępnych danych za pomocą modeli regresji logistycznej można uznać za interesujące.

Literatura

Agresti A., *Categorical Data Analysis,* Wiley, New York 1990.

Cramer J.S., *Logit Models from Econometrics and Other Fields*, University Press, Cambridge 2003.

Jajuga K., *Modele z dyskretną zmienną objaśnianą*, [w:] *Estymacja modeli ekonometrycznych*, red. S. Bartosiewicz, PWE, Warszawa 1990.
Mazurek E., *Analiza danych jakościowych*, [w:] *Statystyczne metody analizy danych*, red. W. Ostasiewicz, AE, Wrocław 1999.
Rodriguez G., *Generalized Linear Models*, http://data.princeton.edu/wws509/notes/ /default.htm.
Stan zdrowia ludności Polski w 1996 r., seria: Informacje i opracowania statystyczne, GUS, Warszawa 1997.
Stan zdrowia, postawy i zachowania zdrowotne mieszkańców Torunia. Raport końcowy z badań wykonanych na zlecenie Wydziału Zdrowia i Pomocy Społecznej Urzędu Miasta Torunia, Katedra Medycyny Społecznej i Zapobiegawczej Akademii Medycznej. Łódź 2001, http://www.um.torun.pl/~gzm/zdrowie/Raport_zdrowie.html.

APPLICATION OF LOGISTIC REGRESSION TO ANALYSIS OF PERCEIVED HEALTH MULTINOMIAL DATA

Summary

The paper presents application of cumulative logit model to the analysis of multinomial data received from state of health research of Torun inhabitants. Logistic regression models have been used to test the dependence of perceived health assessment on demographic characteristics like age, income and education. Proportional odds model with equal slopes for each category were compared with null independence model and full model with regression coefficients estimated for each category. The following statistical relationships between perceived health and demographic characteristics of respondents were shown in the paper:

- the older respondent the poorer perceived health
- the more earning respondent the better perceived health,
- the longer education the better perceived health.