



WYBRANE METODY ANALIZY DANYCH JAKOŚCIOWYCH NA PRZYKŁADZIE WYNIKÓW BADAŃ KARDIOLOGICZNYCH

*Jerzy A. Moczko, Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu,
Katedra Informatyki i Statystyki*

Wprowadzenie

Pomimo dynamicznego rozwoju inżynierii biomedycznej i związanego z tym doskonalenia technik pomiarowych, w badaniach medycznych obok zmiennych ilościowych wielką rolę odgrywają nadal pomiary jakościowe. Charakteryzuje je, podobnie jak zmienne ilościowe, silna zmienność między- i wewnątrzsobnicza. Dla wyciągnięcia wiarygodnych wniosków wymagane jest zatem użycie odpowiednich technik analitycznych. O ile w przypadku pomiarów przeprowadzonych w skali ilorazowej i interwałowej (przedziałowej) znane i powszechnie stosowane jest szerokie spektrum technik badających występowanie istotnych różnic lub też oszacowujących siłę i kształt związków między zmiennymi, o tyle analiza wielkości pomierzonych w skalach porządkowej i nominalnej jest przeprowadzana najczęściej w mało wyszukany sposób. W większości przypadków eksperymentatorzy ograniczają się do konstruowania i prostej interpretacji tablic wielodzielczych, uzasadniając to podejście słabością użytej skali pomiarowej wynikającej z braku jednoznacznie zdefiniowanej odległości między wartościami eksperymentalnymi. Pamiętając, iż zmienne jakościowe znajdują zastosowanie nie tylko w medycynie, lecz również w biologii, psychologii i socjologii i wielu innych dziedzinach nauki, warto przyglądnąć się metodom poszerzonej analizy tych wielkości. Wspomniane skale pomiarowe dopuszczają stosowanie zaawansowanych narzędzi analitycznych jedno- i wielowymiarowych. Przy prawidłowym zebraniu danych, użyciu adekwatnych technik statystycznych i spełnieniu wszystkich wymaganych założeń można uzyskać niezwykle precyzyjne wyniki i na ich podstawie wspomagać rozmaite dziedziny wiedzy. Celem niniejszej prezentacji jest próba wskazania bardziej zaawansowanych technik analitycznych umożliwiających wyciągnięcie z badanego materiału bardziej precyzyjnych wyników.

Opis problemu badawczego

W celu praktycznej ilustracji omawianych technik wykorzystano fragment jednej z bardziej znanych kardiologicznych baz danych – WCGS (*Western Collaborative Group Study*). Projekt WCGS rozpoczęto w roku 1960 w oparciu o 3524 ochotników zatrudnionych



w 11 zakładach pracy w Kalifornii, u których w badaniu wstępnym nie wykryto żadnych oznak chorób układu sercowo-naczyniowego. Jest to typowe badanie obserwacyjne (nierandomizowane). W badaniu uwzględniono m.in. dane socjo-demograficzne (wiek, wykształcenie, zawód, dochód, status małżeński), fizjologiczne (wzrost, masa ciała, ciśnienie krwi, ocenę EKG, wyniki badań okulistycznych), biochemiczne (frakcje lipoprotein, poziom glukozy), dane dotyczące zachowań (typ psychologiczny, palenie tytoniu, konsumpcja alkoholu, aktywność fizyczna), historię występowania chorób układu krążenia w rodzinie). Badania prowadzono przez 25 lat dla potrzeb analizy zachorowalności i śmiertelności spowodowanej przyczynami sercowo-naczyniowymi.

Konstrukcja bazy danych oraz analiza wstępna

Fragment bazy danych WCGS jest dostępny w Internecie m.in. jako materiał ilustracyjny do wspaniałego opracowania Vittinghoffa, Gliddena, Shiboskiego i McCullocha [1] pod adresem <http://www.biostat.ucsf.edu/vgsm/data.html>. Zawiera on 36 zmiennych pomierzonych w rozmaitych skalach pomiarowych dla 3154 przypadków. Bazę importowaliśmy do pakietu *STATISTICA Data Miner 8* i wybraliśmy z niej losowo podzbiór 6 zmiennych oraz 569 przypadków przydatnych do przedstawienia analizy postawionego problemu. Podzieliliśmy je na dwie grupy: 3 zmienne niezależne – predyktory (wiek w chwili włączenia pacjenta do badania, występowanie *arcus senilis*: zmętnienie na krańcach rogówki związane z hipercholesterolemią i palenie tytoniu) oraz zmienna zależna binarna CHD69 (wystąpienie choroby wieńcowej w okresie 25 lat badania). Wiek przedstawiono w trzech możliwych skalach (interwałowej – AGE, porządkowej – AGECE oraz nominalnej – BAGE50: wiek powyżej 50 lat) w celu ilustracji wpływu zmiany skali pomiarowej na uzyskiwane wyniki. Będziemy próbowali odpowiedzieć na pytanie, które z badanych czynników mają obserwowalny i istotny statystycznie wpływ na wystąpienie w badanej grupie 569 pacjentów choroby wieńcowej.

	1	2	3	4	5	6
	age	agec	bage_50	arcus	smoke	chd69
1	50	2	1	1	1	0
2	51	3	1	0	1	0
3	59	4	1	1	0	0
4	51	3	1	1	0	0
5	44	1	0	0	0	0
6	47	2	0	0	1	0
7	40	0	0	0	0	0
8	41	1	0	0	1	0
9	50	2	1	1	0	0
10	43	1	0	0	1	0
11	59	4	1	0	1	0
12	54	3	1	0	0	0
13	48	2	0	0	1	0
14	49	2	0	0	1	0
15	55	3	1	1	0	0
16	44	1	0	0	0	0
17	56	4	1	0	1	0
18	42	1	0	0	0	0
19	44	1	0	0	0	0
20	45	1	0	1	0	0
21	42	1	0	0	0	0

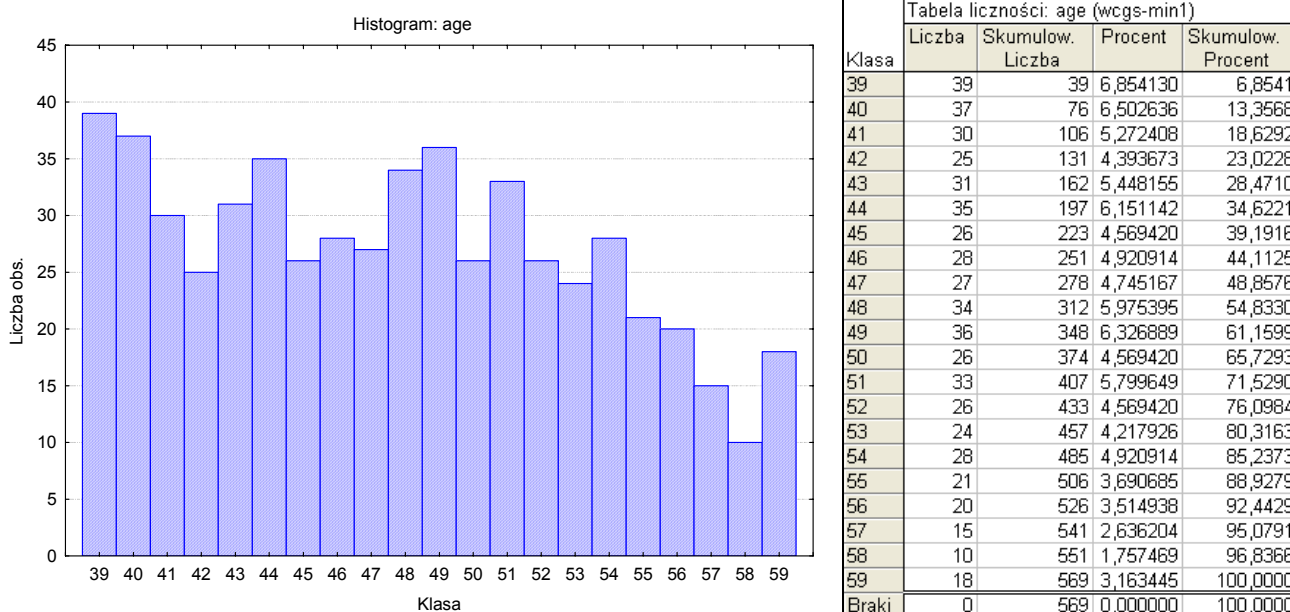
Rys. 1. Fragment bazy danych użytej do ilustracji technik analizy danych kategoryalnych.



Ograniczenie liczby zmiennych i dobór przypadków ma na celu ułatwienie Czytelnikowi zrozumienia użytych procedur statystycznych i w żadnym przypadku przytaczanych wyników obliczeń nie należy cytować jako prawdziwe wnioski medyczne (aczkolwiek zdecydowana większość wyników pokrywa się z oryginalnymi wynikami projektu WCGS). Fragment analizowanej bazy przedstawiony jest na rys. 1.

Pierwszą narzucającą się automatycznie metodą analizy danych jest zestawienie kategorii, jakie występują dla każdej z analizowanych zmiennych. Oczywiście nie daje to odpowiedzi na postawione pytanie dotyczące znalezienia czynników decydujących o wystąpieniu choroby wieńcowej, lecz pozwala się zorientować ogólnie w strukturze badanego materiału i zaprojektować dalsze, bardziej zaawansowane analizy.

Rozpocznijmy analizę od bliższego przyjrzenia się zmiennej *wiek*. Jak wspomnieliśmy, zakodowano go w trzech skalach pomiarowych. Rys. 2 przedstawia histogram i liczebności poszczególnych kategorii wiekowych. W cytowanej uprzednio pracy Vittinghoffa i in. [1] przyjęto przy przejściu do skali porządkowej podział na 5 grup wiekowych, począwszy od wieku 35 lat. Jak łatwo zauważyć, przyglądając się rys. 2, podział taki nie jest optymalny ze względu na całkowity brak pomiarów w kategoriach 35-38 lat oraz 60 lat. Można oczywiście, wykorzystując dostępne opcje programu *STATISTICA*, dokonać optymalizacji histogramu, ale przyjęliśmy wersję podziału przyjętą w pracy [1], by umożliwić Czytelnikowi ewentualne spójne porównywanie wyników.



Rys. 2. Histogram i tabela licznosci zmiennej *wiek* wyrażonej na skali interwałowej.

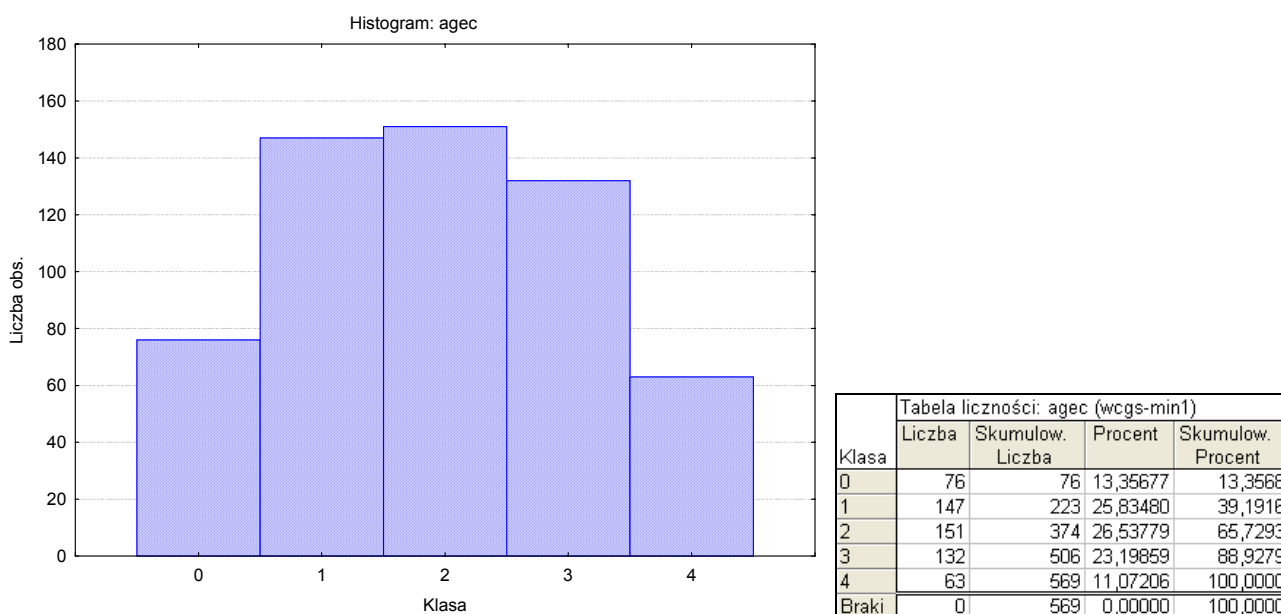
Przyjrzyjmy się teraz dla przykładu dwom tablicom wynikowym: rozkładowi danych dotyczących palenia tytoniu (rys. 3) oraz grup wiekowych (rys. 4).



Tabela liczności: smoke (wcgs-min1)					Tabela liczności: smoke (wcgs-min1)		
Klasa	Liczba	Skumulow. Liczba	Procent	Skumulow. Procent	Klasa	Liczba	Procent
0	255	255	44,81547	44,8155	0	255	44,81547
1	314	569	55,18453	100,0000	1	314	55,18453
Braki	0	569	0,00000	100,0000	Braki	0	0,00000

Rys. 3. Nieprawidłowo (strona lewa) i prawidłowo prezentowana tablica rozkładu osób palących i niepalących w badanym materiale (kod 0 – niepalący, kod 1 – palący).

Jak łatwo zauważyć, niejednorodność występowania osób palących w badanym materiale jest stosunkowo słaba, natomiast obserwujemy istotne niejednorodności w strukturze grup wiekowych. Zasadnicza różnica między obiema tabelami polega na tym, że w tabeli na rys. 3 analizujemy zmienną nominalną, natomiast w tabeli na rys. 4 – zmienną porządkową. Aczkolwiek uwaga ta może wydawać się trywialną, próba interpretacji przytaczanych wyników w kolumnach „skumulowana liczba” i „skumulowany procent” w tabeli z rys. 2 jest bezsensowna (niestety mimo wszystko pojawiają się prace z takimi „rodzinkami”). Wiersze lub kolumny, których nie chcemy na ilustracji, wystarczy po kliknięciu na ich nagłówku prawym klawiszem myszki usunąć.



Rys. 4. Tablica rozkładu grup wiekowych w badanym materiale (kod 0: 35-40 lat, kod 1: 41-45 lat, kod 2: 46-50 lat, kod 3: 51-55 lat, kod 4: 56-60 lat).

Na podstawie zamieszczonego powyżej histogramu i tabeli liczności możemy stwierdzić, że dwie pierwsze kategorie wiekowe są nieco słabiej reprezentowane niż pozostałe w porównaniu do rozkładu wieku przedstawionego na rys. 2.

Podstawowe techniki analizy danych kategoryalnych

Najprostszym i najczęściej stosowanym sposobem określenia wpływu pojedynczego predyktora na wartość dychotomicznej zmiennej zależnej jest wykorzystanie tablic



wielodzielczych. Na rys. 5 przedstawiono tabelę dwudzielczą, ilustrującą występowanie choroby wieńcowej oraz wieku badanych wyrażonego w skali porządkowej. Tabela została wzbogacona o procenty wyliczone w układzie wierszowym, kolumnowym oraz w stosunku do całkowitej liczebności badanej próby.

Podsumowująca tabela dwudzielcza: częstości obserwowane						
Liczebność oznacz. komórek > 10						
chd69	agec 0	agec 1	agec 2	agec 3	agec 4	Wiersz Razem
0	45	92	81	67	27	312
% z kolu	59,21%	62,59%	53,64%	50,76%	42,86%	
% z wier	14,42%	29,49%	25,96%	21,47%	8,65%	
% z cał	7,91%	16,17%	14,24%	11,78%	4,75%	54,83%
1	31	55	70	65	36	257
% z kolu	40,79%	37,41%	46,36%	49,24%	57,14%	
% z wier	12,06%	21,40%	27,24%	25,29%	14,01%	
% z cał	5,45%	9,67%	12,30%	11,42%	6,33%	45,17%
Ogół	76	147	151	132	63	569
% z cał	13,36%	25,83%	26,54%	23,20%	11,07%	100,00%

statystyka	Statystyka: chd69(2) x agec(5)		
	Chi-kwadr.	df	p
Chi kwadrat Pearsona	8,774857	df=4	p=,06699
Chi*2 NW	8,805382	df=4	p=,06616

Rys. 5. Tabela dwudzielcza ilustrująca wpływ wieku badanego wyrażonego w skali porządkowej na wystąpienie choroby wieńcowej.

Na pierwszy rzut oka widoczny jest fakt braku zależności badanych zmiennych, co potwierdzają nieistotne statystycznie wartości testów Chi-kwadrat wg Pearsona ($p=0,06699$) oraz Chi-kwadrat największej wiarygodności ($p=0,06616$).

Stoi to w sprzeczności ze „zdrowym rozsądkiem” i udowodnionymi faktami medycznymi. Rzeczywiście, jeżeli porównamy przy użyciu parametrycznego testu *t*-Studenta dla zmiennych niepowiązanych średni wiek badanych w grupie osób z chorobą wieńcową i bez tej choroby (rys. 6), obserwujemy istotną statystycznie różnicę wynoszącą około 1,5 roku. Możemy zatem podejrzewać, że przejście w opisie wieku ze skali interwałowej (AGE) do skali porządkowej (AGEC) spowodowało tak dużą stratę informacji, iż nie możemy wykazać istniejącej w rzeczywistości różnicy wieku. To tłumaczenie jest jednakże fałszywe.

Testy t; Grupaująca: chd69 (wcgs-min1)														
Grupa 1: 0														
Grupa 2: 1														
Zmienna	Średnia 0	Średnia 1	t	df	p	N ważnyc 0	N ważnych 1	Odch.std 0	Odch.std 1	iloraz F	p	Levene'a F(1,df)	df	p
age	47,13462	48,49027	-2,81946	567	0,004979	312	257	5,629603	5,801477	1,061993	0,611565	0,053639	567	0,816931

Rys. 6. Porównanie średniego wieku badanych między grupami osób bez choroby wieńcowej (średnia 0) i z występującą chorobą wieńcową (średnia 1).

Spróbujmy bowiem zbadać tę różnicę, wykorzystując jeszcze słabszy sposób pomiaru – skalę nominalną (BAGE50) (rys. 7.). Dla osób młodszych (wiek do 50 lat włącznie) przyjęto kod 0, dla starszych – kod 1. Wyniki zestawione w tabeli dwudzielczej utworzonej dla zmiennych CHD69 oraz BAGE50 pozwalają na wyciągnięcie wniosku, że między badanymi zmiennymi istnieje istotny statystycznie związek ($p=0,03526$). Widać zatem, że nasze poprzednie tłumaczenie powodu braku związku między zmiennymi CHD69 a AGEc nie było słuszne. Prawdziwą przyczyną była nie utrata informacji przy przejściu między skalami pomiarowymi, ale użycie niewłaściwego testu (o zbyt niskiej mocy) do analizy przekodowanych danych.



Podsumowująca tabela dwudzielcza: c			
Liczność oznacz. komórek > 10			
chd69	bage_50 0	bage_50 1	Wiersz Razem
0	203	109	312
% z kolu	58,33%	49,32%	
% z wier	65,06%	34,94%	
% z cał	35,68%	19,16%	54,83%
1	145	112	257
% z kolu	41,67%	50,68%	
% z wier	56,42%	43,58%	
% z cał	25,48%	19,68%	45,17%
Ogół	348	221	569
% z cał	61,16%	38,84%	100,00%

Statystyka: chd69(2) x bage_50(2)			
statystyka	Chi-kwadr.	df	p
Chi kwadrat Pearsona	4,432460	df=1	p=,03526
Chi ² Nw	4,427996	df=1	p=,03536
Chi ² Yatesa	4,076046	df=1	p=,04350
dokt. Fishera, 1-stronny			-----
2-stronny			
Chi ² McNemara (A/D)	25,71428	df=1	p=,00000
(B/C)	4,822834	df=1	p=,02809

Rys. 7. Analiza współwystępowania choroby wieńcowej (CHD69) i wieku kodowanego w skali nominalnej (BAGE50).

Przyjrzyjmy się ponownie zmiennej AGECE. Spróbujmy spojrzeć na tabelę kontyngencji z rys. 5 po obróceniu jej o 90° w prawo (omijając oczywiście wszystkie wyniki procentowe). Tabela przyjmie postać przedstawioną na rys. 8. Jej struktura ułatwi nam zrozumienie faktu, dlaczego do jej analizy powinno się użyć np. testu Manna–Whitney’a w miejscu testu chi-kwadrat. Nasza zmienna pomiarowa AGECE posiada naturalne uporządkowanie wartości i dlatego powinna być interpretowana w sposób następujący: w kolumnie CHD69=0 występuje 45 pomiarów o wartości 0, 92 pomiary o wartości 1, 81 pomiarów o wartości 2 itd. Łącznie w kolumnie tej mamy 312 pomiarów. Podobnie interpretujemy kolumnę CHD69=1, w której łącznie mamy 257 pomiarów. Ponieważ pomiary w obu kolumnach są reprezentowane w skali porządkowej, do ich porównania użyjemy np. testu Manna-Whitneya. Pamiętajmy jednak, że w programie *STATISTICA* wymagane jest, aby dane zostały zorganizowane w podobny sposób, jak dla testu *t*-Studenta dla prób niezależnych (zmienna wartości pomiarowych i zmienna kodowa, grupująca).

Tabela licznosci (wcgs-min1)			
Liczność oznacz. komórek > 10			
(Nie oznaczono sum brzegowych)			
agec	chd69 0	chd69 1	Wiersz Razem
0	45	31	76
1	92	55	147
2	81	70	151
3	67	65	132
4	27	36	63
Ogół grp	312	257	569

Rys. 8. Przekształcona tabela dwudzielcza z rys. 5.

W wyniku użycia do zmiennej AGECE testu Manna–Whitney’a otrzymujemy istotną statystycznie różnicę ($p=0,006946$) wieku między grupami badanych z chorobą wieńcową i bez niej. Możemy zatem podsumować, że niezależnie od sposobu wyboru skali pomiarowej do określenia wieku pacjenta w badanej populacji zawsze jesteśmy w stanie wykazać związek między wiekiem a występowaniem choroby wieńcowej.

Na zakończenie warto jeszcze opisać jeden stosunkowo często popełniany błąd – próbę użycia testu Manna-Whitney’a bezpośrednio do wyników zawartych w tabeli dwudzielczej z rys. 8. Jeżeli zgodnie z wymaganiami organizacji danych w pakiecie *STATISTICA* wprowadzimy zmienną grupującą CHD, lecz nie wykonamy opisanego przekodowania danych



wieku, to uzyskamy błędny wynik (rys. 9). Spowodowane jest to faktem, iż postępując w ten sposób porównujemy rozkłady liczebności grup wiekowych, a nie rozkłady wartości pomiarowych opisujących wiek.

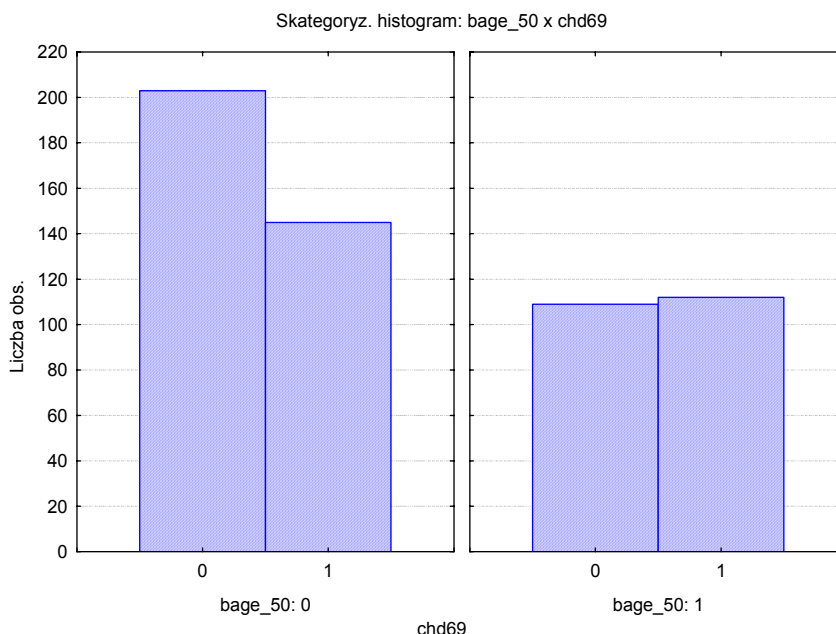
	1 AGEC	2 CHD
1	45	0
2	92	0
3	81	0
4	67	0
5	27	0
6	31	1
7	55	1
8	70	1
9	65	1
10	36	1

Test U Manna-Whitneya (Arkuszt14) Względem zmiennej: CHD Zaznaczone wyniki są istotne z $p < ,05000$										
Zmienna	Sum.rang Grupa 1	Sum.rang Grupa 2	U	Z	poziom p	Z popraw.	poziom p	N ważn. Grupa 1	N ważn. Grupa 2	2*1 str. dokł. p
AGEC	31,00000	24,00000	9,000000	0,731126	0,464703	0,731126	0,464703	5	5	0,547619

Rys. 9. Błędny sposób kodowania danych z rys. 5.

Analizy wielowymiarowe – podejście klasyczne

Przyjrzyjmy się teraz sytuacjom, w których będziemy usiłowali zbadać jednoczesny wpływ na jakościową zmienną zależną dwóch lub większej liczby predyktorów. Można oczywiście tworzyć wiele możliwych kombinacji analiz wieloczynnikowych, generując w ten sposób obok modeli prostych (wpływ pojedynczego predyktora na zmienną decyzyjną) modele złożone, zawierające oprócz efektów prostych, pochodzących z modeli prostych, efekty interakcyjne.



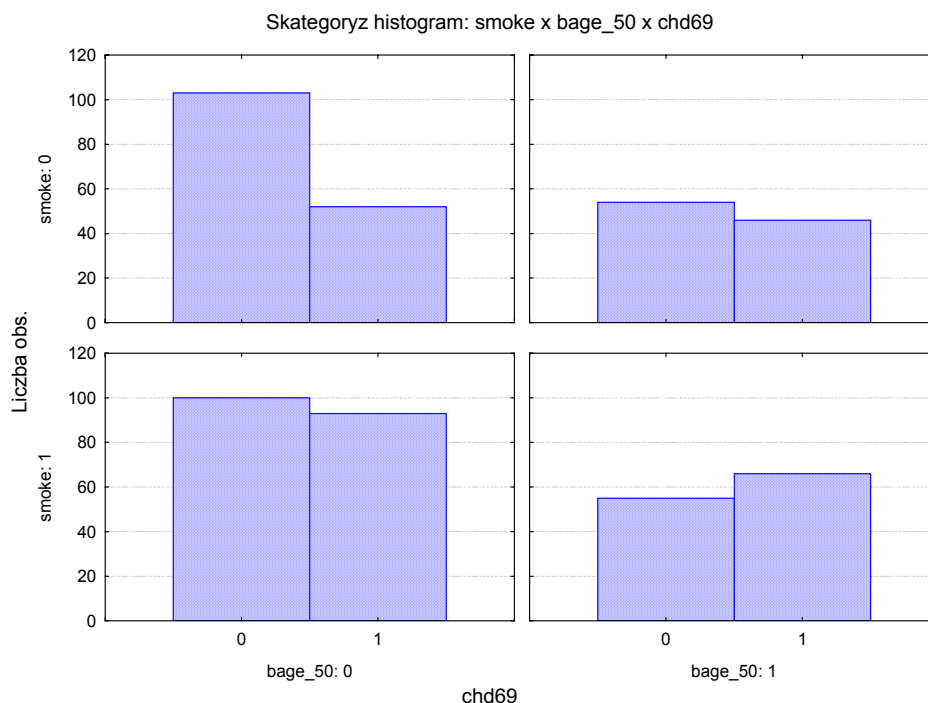
Rys. 10. Ilustracja graficzna wpływu wieku na występowanie w badanej populacji choroby wieńcowej.

Spróbujmy na przykład przyjrzeć się, czy i w jaki sposób palenie tytoniu modyfikuje zależność występowania choroby wieńcowej w zależności od wieku. Przyjrzyjmy się najpierw wykazanej w poprzednim rozdziale zależności CHD69 od wieku BAGE50, wykorzystując histogramy skategoryzowane. Jak widać na rys. 9 (powyżej), początkowa przewaga liczby osób bez choroby wieńcowej (CHD69=0) nad liczbą osób mających tę jednostkę chorobową (CHD69=1) maleje z wiekiem, a nawet w grupie osób starszych obserwujemy lekką przewagę osób z chorobą wieńcową.

Przeprowadźmy teraz taką samą analizę, lecz uwzględniając dodatkowo informację o paleniu tytoniu. Jak widać w górnym wierszu rys. 11, w grupie niepalących (SMOKE=0) obserwujemy taką samą tendencję jak poprzednio (spadek proporcji liczby osób bez CHD do liczby osób z CHD ze wzrostem wieku). Jest ona nadal istotna statystycznie ($p=0,04597$). W dolnym wierszu, ilustrującym grupę palaczy, tendencja ta staje się nieistotna ($p=0,27270$). Można to interpretować w dwojaki sposób:

1. Palenie tytoniu „konserwuje” badanych tak, że starzenie nie wpływa na wzrost proporcji liczby z CHD do liczby osób bez CHD.
2. Palenie tytoniu wywołało dużą liczbę zachorowań na chorobę wieńcową już w młodszym wieku i ta wysoka frakcja zachorowań pozostaje niezmienna z wiekiem.

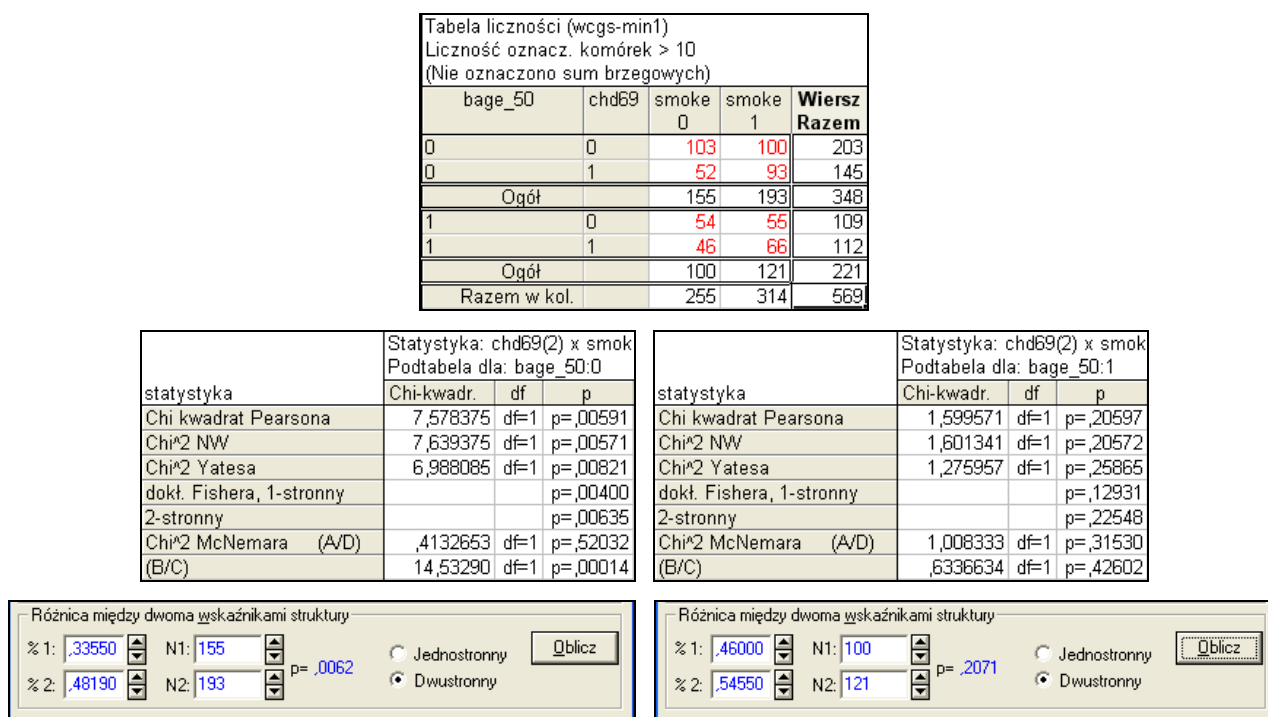
Aby odpowiedzieć na pytanie, która z odpowiedzi jest właściwa, porównajmy frakcję osób z CHD wśród palaczy i niepalących w obrębie osób młodszych. Możemy tego dokonać albo stosując ponownie opcję tablic wielodzzielczych, albo przy użyciu testu różnicy między dwoma wskaźnikami struktury.



Rys. 11. Ilustracja graficzna wpływu wieku na występowanie w badanej populacji choroby wieńcowej przy jednoczesnym uwzględnieniu informacji o paleniu tytoniu.



Wyniki obu typów analiz są zgodne i wskazują na istotną statystycznie przewagę frakcji osób z CHD w grupie palących wśród grupy młodszych badanych. Zamieszczono je na rys. 12.



Rys. 12. Wyniki porównania między palącymi i niepalącymi frakcji osób z CHD w grupie młodszych przy użyciu testu Chi-kwadrat (górny wiersz) i testu różnicy między wskaźnikami struktury (dolny wiersz).

Analizy wielowymiarowe – technika log-liniowa

Jak widać, analiza nawet stosunkowo prostego modelu z dwoma binarnymi predyktorami wymaga dużego nakładu pracy. Łatwo sobie wyobrazić, jak złożonym stanie się problem, w którym będziemy uwzględniać większą liczbę predyktorów o rozkładzie wielomianowym. Przede wszystkim należy sobie zdać sprawę z faktu, że wzrost liczby czynników w modelu powoduje zwiększenie liczby podziałów wyjściowej liczby przypadków, co doprowadzi do powstania macierzy rzadkich opierających się tradycyjnym metodom analizy statystycznej. W literaturze statystycznej problem ten zwany jest „przekleństwem wymiarowości” (*curse of dimensionality*). Wyniki takich analiz będą niewiarygodne. W miejsce asymptotycznych oszacowań prawdopodobieństwa będziemy zmuszeni stosować obliczenia dokładne, takie jak dokładne algorytmy permutacyjne lub techniki Monte Carlo [2]. Wręcz niezbędnymi okażą się te techniki w przypadku macierzy rzadkich (tablice kontyngencji z przeważającą liczbą komórek o wartości zero), zestawów danych z małymi liczebnościami komórek lub obsadzeniami silnie niezbalansowanymi (w niektórych komórkach bardzo duże liczebności, w pozostałych małe lub wręcz zerowe). Z drugiej strony liczba możliwych kombinacji porównań par metodą testu różnicy frakcji



będzie rosła z liczbą włączanych zmiennych i liczbą kategorii charakteryzujących daną zmienną, dlatego też w praktyce nie używa się tabel wielodzzielczych dla więcej niż trzech zmiennych jednocześnie.

W sytuacji, gdy z jakichkolwiek przyczyn musimy użyć do analizy jednocześnie wielu czynników, jedną z najbardziej efektywnych technik okazuje się analiza log-liniowa [3]. Buduje ona i testuje modele przedstawiające istniejące w badanej populacji zależności i interakcje na poziomie zmiennych jakościowych. Można ją traktować jako pewien model regresji oparty wyłącznie na zmiennych jakościowych. Zakładamy, że wszelkie istotne odstępstwa liczebności obserwowanych od liczebności oczekiwanych są wskaźnikami istnienia zależności między badanymi zmiennymi. Innymi słowy, model log-liniowy przeprowadza weryfikację hipotezy zerowej zakładającej brak jakichkolwiek interakcji dwóch lub większej liczby zmiennych jakościowych. Dla przykładu zbudujemy model badający interakcje jednocześnie czterech zmiennych pomierzonych w binarnej skali nominalnej (rys. 13). Spróbujmy teraz dokonać analizy otrzymanych wyników.

Efekt	Testy związku brzegowego i cząstkowego (wcdgs-min1)				
	Stopnie swobody	Zw. cząst chi-kwad	Zw. cząst p	Z. brzeg. chi-kwad	Z. brzeg. p
1	1	28,28311	0,000000	28,28311	0,000000
2	1	50,41245	0,000000	50,41245	0,000000
3	1	6,06460	0,013792	6,06460	0,013792
4	1	5,65975	0,017359	5,65975	0,017359
12	1	6,43051	0,011218	7,07560	0,007814
13	1	0,51231	0,474142	0,06582	0,797525
14	1	3,69414	0,054604	4,32282	0,037605
23	1	3,78435	0,051734	4,52115	0,033478
24	1	3,03326	0,081574	4,84515	0,027724
34	1	7,73035	0,005430	8,45071	0,003649
123	1	0,41220	0,520857	0,74023	0,389586
124	1	1,91467	0,166446	2,25339	0,133322
134	1	0,45859	0,498284	0,73779	0,390369
234	1	0,11421	0,735398	0,16370	0,685773

Najlepszy początk.: Chi-kwadrat= 3,5046 df = 5 p = ,6227
21,31,32,41,42,43

Najlepszy: Chi-kwadrat= 7,1005 df = 7 p = ,4185
21,43,32,41

Analizowana tabela:

(1)	(2)	(3)	(4)
bage_50	arcus	smoke	chd69
2	x 2	x 2	x 2

Min. liczebność komórki: 16, Maksimum: 81, Suma: 567,

Testowany model: 21,43,32,41

Delta: ,5000 ; Maks. iteracji: 50 ; Kryt. zbieżności: ,0100
Osiągnięto zbieżność po # iteracjach: 3

	df	p
Chi-kwadrat najw. wiaryg.:	7,1005	7,41850
Chi-kwad. Pearsona:	7,2045	7,40791

Tabela: bage_50(2) * arcus(2) * smoke(2) * chd69(2) Model: 21,43,32,41			
Test	Chi-kw.	df	p
Chi-kwad. najw. wiaryg.	7,100464	7	0,418495
Chi-kwad. Pearsona	7,204460	7	0,407907

Rys. 13. Wyniki analizy log-liniowej.



W pierwszym kroku analizujemy wartości zależności cząstkowej odpowiadającej na pytanie, czy dany element modelu jest istotny, gdy wszystkie pozostałe efekty tego samego rzędu są już wbudowane do modelu. Jak widać, istotne są wszystkie efekty proste (bez interakcji) oraz następujące efekty drugiego rzędu: 12 i 34 (BAGE50*ARCUS oraz SMOKE*CHD69). Żaden element interakcji trzeciego rzędu nie jest istotny. Kolejno przechodzimy do analizy wartości zależności brzegowej. Mówi ona, czy dana interakcja określonego rzędu odgrywa rolę w modelu, w którym nie uwzględniono jeszcze żadnych innych interakcji tego samego rzędu. Jak widać, istotne są wszystkie interakcje 2 rzędu za wyjątkiem interakcji 13 (BAGE50*SMOKE). Oznacza to, że efekt tej interakcji został już w modelu w pełni wytłumaczony przez włączenie pozostałych istotnych efektów. Ponieważ efekty interakcji rzędu k obejmują automatycznie efekty interakcji rzędu $k-1$, możemy ostatecznie jako istotny dla nas przyjąć model log-liniowy z efektami: 21,32,41,42,43 (kolejność wystąpienia zmiennej w interakcji nie odgrywa roli). Zastosowanie automatycznej procedury doboru zmiennych doprowadza nas do niemal identycznego doboru modelu (zostaje jedynie wyeliminowana dodatkowo interakcja 42, dla której zależność cząstkowa nie jest istotna na poziomie $p=0,08157375$).

Testy oceniające dobroć dopasowania wytworzonego ostatecznie modelu (Chi-kwadrat Pearsona i największej wiarygodności) wskazują, że wyjaśnia on w zadowalający sposób liczebności występujące w naszej czterowymiarowej tabeli kontyngencji. Ponieważ postawiony przez nas problem dotyczy wyboru, które zmienne wpływają w istotny sposób na zmienną zależną CHD69 (w naszym modelu zmienna nr 4) dochodzimy ostatecznie do wniosku, że są to predyktory BAGE50 (zmienna 1) oraz SMOKE (zmienna 3). Ostatnim krokiem analizy jest przeglądnięcie tabel obserwowanych, dopasowanych, brzegowych, reszt i reszt standaryzowanych.

Analizy wielowymiarowe – regresja logistyczna

Omówiona w poprzednim rozdziale technika analizy log-liniowej w zasadniczy sposób ułatwiła analizę wielowymiarowych tablic kontyngencji w porównaniu z podejściem klasycznym. Nadal jednakże analiza sprawi nam wiele problemów w sytuacji, gdy predyktory będą mogły przyjmować wiele wartości, a w szczególności gdy będą zmiennymi ciągłymi [1, 4, 5, 6]. W praktycznych analizach medycznych sytuacja taka pojawia się powszechnie. Predyktory binarne, takie jak: płeć, palenie tytoniu, występowanie jakiegoś typu objawy, współlistnieją z predyktorami o rozkładzie wielomianowym (stopień zaawansowania choroby, grupa wiekowa, efekt leczenia, genotyp) i z predyktorami ciągłymi (poziomy glukozy, trójglicerydów, cholesterolu, BMI). W analizowanej przez nas bazie wiek badanego zakodowaliśmy w trojaki sposób: jako zmienną w skali interwałowej (AGE), porządkowej (AGEC) oraz nominalnej (BAGE50). Metody dotychczas omówione bez problemu radzą sobie z dwoma ostatnimi typami zmiennych. Może pojawić się jednak pytanie związane z arbitralnością podziału kategoryjnego – dlaczego przyjęto akurat takie granice przedziałów, a nie inne. Może się okazać, że przy jednym podziale otrzymamy istotne zależności między zmiennymi, przy innym nieistotne. Co więcej, kierunek zależności może ulec odwróceniu (paradoks Simpsona). W niektórych sytuacjach byłoby zatem korzystne



uniknięcie przekodowywania danych. Aby móc jednocześnie uwzględnić predyktory pomierzone w skali interwałowej, należy użyć analizy logistycznej. Metoda ta w dużym stopniu przypomina wieloraką regresję liniową. Dwie podstawowe różnice to binarność zmiennej zależnej (zamiast zmiennej ciągłej), a co z tym się wiąże inny sposób estymacji parametrów modelu regresyjnego (metoda największej wiarygodności w miejscu metody najmniejszych kwadratów).

Model: Regr. logistyczna (logit) N zer: 312 jedynek: 255 (wcgs-min1) Zmn. zal.: chd69 Strata: Największe prawd. bł. średnk.w. skal. Całkowita strata: 382,04947684 Chi2(3)=16,190 p=,00104				
N=567	Stała B0	bage_50	arcus	smoke
Ocena	-0,7165503	0,339971	0,3136564	0,4837745
Błąd standard.	0,157033	0,1762064	0,1800596	0,1733852
t(563)	-4,563056	1,929391	1,741959	2,790172
poziom p	0,000006193032	0,0541845	0,08206175	0,005446526
-95%CL	-1,024992	-0,00613116	-0,04001425	0,1432136
+95%CL	-0,4081082	0,6860731	0,667327	0,8243353
Chi-kwadrat Walda	20,82147	3,722551	3,03442	7,78506
poziom p	0,000005060166	0,05369113	0,08152524	0,005271119
iloraz szans z.jedn.	0,4884343	1,404907	1,368419	1,622186
-95%CL	0,3587992	0,9938876	0,9607757	1,153976
+95%CL	0,6649069	1,985902	1,949021	2,280365
iloraz szans zakr.		1,404907	1,368419	1,622186
-95%CL		0,9938876	0,9607757	1,153976
+95%CL		1,985902	1,949021	2,280365

Model: (wcgs-min1) Zmn. zal.: chd69			
	Obserw.	Przewidyw.	Reszty
1	0,000000	0,603687	-0,603687
2	0,000000	0,526773	-0,526773
3	0,000000	0,484274	-0,484274
4	0,000000	0,484274	-0,484274
5	0,000000	0,328153	-0,328153
6	0,000000	0,442067	-0,442067
7	0,000000	0,328153	-0,328153
8	0,000000	0,442067	-0,442067
9	0,000000	0,484274	-0,484274
10	0,000000	0,442067	-0,442067
11	0,000000	0,526773	-0,526773
12	0,000000	0,406952	-0,406952
13	0,000000	0,442067	-0,442067
14	0,000000	0,442067	-0,442067

Klasyfikacja przypadków (wcgs-min1) Il. szans: 1,8393 % poprawnych: 58,38%			
Obserw.	Przew.	Przew.	Procent
	0,000000	1,000000	Popraw.
0,000000	227	85	72,75641
1,000000	151	104	40,78431

Rys. 14. Wyniki analizy logistycznej przy użyciu zmiennej wiek wyrażonej w skali nominalnej.

Warto wspomnieć, że istnieją również modele wielopoziomowej regresji logistycznej (*polytomous regression*) pozwalające na konstruowanie modeli logitowych wielomianowych dla nieuporządkowanej lub uporządkowanej odpowiedzi kategoryjnej, jednakże ich omówienie jest poza zakresem tematyki niniejszej pracy. Spróbujmy na początku przyjrzeć się wynikom analizy logistycznej, dokonując identycznego wyboru zmiennych jak w przypadku przeprowadzonej w poprzednim rozdziale analizy log-liniowej. Jedyna różnica polega na tym, iż w analizie logistycznej w sposób jawny definiujemy zmienną zależną, czego nie czyniliśmy w analizie log-liniowej. Przeanalizujemy wyniki tej analizy zamieszczone na rys. 14. Jak widać, utworzony przez nas model różni się w sposób istotny statystycznie od modelu zawierającego wyłącznie wyraz wolny ($p=0,00104$) i dlatego jest sensowne przyjrzenie mu się dokładniej. Spośród włączanych do modelu predyktorów jedynie jeden z nich jest istotny statystycznie (zmienna SMOKE, $p=0,00545$). Zmienna BAGE50, którą do tej pory w każdej analizie wykazywaliśmy jako istotnie powiązaną z występowaniem



choroby wieńcowej, znalazła się praktycznie na granicy decyzyjnej ($p=0,05419196$, dolny 95% przedział ufności $-0,006141921$, górny $0,6860839$). I jakkolwiek trzymając się ściśle reguł przyjmowania lub odrzucania hipotezy zerowej, brak jest dowodu na istotny wpływ zmiennej BAGE50 na zmienną zależną CHD69, to widzimy, że obliczona wartość prawdopodobieństwa leży tuż obok progu decyzyjnego i zmiana pojedynczego przypadku może zmienić naszą decyzję o odrzuceniu BAGE50 jako zmiennej nieistotnej w modelu logistycznym. W takiej sytuacji praktycy pozostawiają zmienną w konstruowanym modelu. Ocenę siły wpływu danej zmiennej uzyskamy, analizując jednostkowy iloraz szans. Jak widać, palenie tytoniu zwiększa ryzyko wystąpienia choroby wieńcowej aż 1,64 razy w stosunku do osób niepalących, przejście do wyższej grupy wiekowej zwiększa ryzyko o około 1,4 razy.

Równanie logitowe przyjmuje więc postać:

$$\text{Logit } P = 0,4837745 * \text{SMOKE} + 0,339971 * \text{BAGE50} - 0,7165503$$

Zbudowany model logistyczny pozwala nam na ocenę jakości klasyfikacji przypadków użytych do budowy modelu oraz oszacowanie ilorazu szans. Jak widać, model ten nie jest wysokiej jakości (poprawna klasyfikacja 58,38% przypadków). Iloraz szans (liczony jako stosunek iloczynu liczby przypadków prawidłowo sklasyfikowanych do liczby przypadków błędnie sklasyfikowanych) nie jest wiele większy od jedności, co oznacza niską jakość modelu. Jeżeli weźmiemy pod uwagę, że dokonaliśmy klasyfikacji typu post-hoc (przypadków użytych do budowy modelu), należy się spodziewać, że zbudowany model nie nadaje się do praktycznych celów predykcyjnych. O ile predykcja braku CHD jest w miarę dobra (około 73% prawidłowych klasyfikacji), o tyle status palenia tytoniu i wiek powyżej 50 lat nie jest wystarczającym prognostykiem wystąpienia choroby wieńcowej (około 41%). Korzystając z opcji *Obserwowane, Przewidywane, Reszty* łatwo możemy zidentyfikować, które przypadki zostały sklasyfikowane błędnie przez nasz model.

Spróbujmy teraz zastąpić zmienną BAGE50 (wiek wyrażony w skali nominalnej) przez zmienną AGECE (wiek wyrażony w skali porządkowej) (rys. 15).

Model: Regr. logistyczna (logit) N zer: 312 jedynek: 255 (wcgs-min1) Zmn. zal.: chd69 Strata: Największe prawd. bł.średnkw.skala. Całkowita strata: 380,55660096 Chi2(3)=19,176 p=.00025				
N=567	Stała B0	agec	arcus	smoke
Ocena	-0,9440069	0,185839	0,2966928	0,4931563
Błąd standard.	0,1998215	0,07214104	0,180603	0,1739967
t(563)	-4,72425	2,576051	1,64279	2,834285
poziom p	0,000002920416	0,01024733	0,1009847	0,004757667
-95%CL	-1,336494	0,04414057	-0,05804516	0,1513942
+95%CL	-0,5515201	0,3275375	0,6514308	0,8349183
Chi-kwadrat Walda	22,31854	6,636042	2,69876	8,033169
poziom p	0,00000231946	0,009998016	0,1004362	0,004595717
iloraz szans z.jedn.	0,3890657	1,204228	1,345402	1,637476
-95%CL	0,2627654	1,045129	0,9436073	1,163455
+95%CL	0,5760735	1,387547	1,918284	2,304626
iloraz szans zakr.		2,102982	1,345402	1,637476
-95%CL		1,193109	0,9436073	1,163455
+95%CL		3,706729	1,918284	2,304626

Klasyfikacja przypadków (wcgs-min1) Il. szans: 2,3291 % poprawnych: 60,85%			
Obszew.	Przew.	Przew.	Procent Popraw.
0,000000	239	73	76,60256
1,000000	149	106	41,56863

Rys. 15. Wyniki analizy logistycznej przy użyciu zmiennej wiek wyrażonej na skali porządkowej.

Podobnie jak poprzednio wszystkie zmienne predykcyjne z wyjątkiem zmiennej ARCUS są istotne statystycznie, a równanie logitowe przyjmuje postać



$$\text{Logit } P = 0,4931563 * \text{SMOKE} + 0,185839 * \text{AGEC} - 0,9440069$$

Wzrosła o około 2% liczba prawidłowo rozpoznanych przypadków, głównie w grupie osób bez CHD. Wartość współczynnika regresji przy zmiennej SMOKE praktycznie nie uległa zmianie, co oznacza, że iloraz szans przy zmianie jednostkowej w obu modelach jest niemal identyczny i wynosi około 1.64. Z oczywistych względów spadła natomiast wartość ilorazu szans przy zmianie jednostkowej dla zmiennej wieku AGECE.

Ostatni eksperyment polega na użyciu wieku zakodowanego w skali interwałowej (AGE) (rys. 16).

Model: Regr. logistyczna (logit) N zer: 312 jedynek: 255 (wcgs-min1)				
Zmn. zal.: chd69 Strata: Największe prawd. bł.średnkw.skal.				
Całkowita strata: 380,22435695 Chi2(3)=19,840 p=.00018				
N=567	Stała BD	age	arcus	smoke
Ocena	-2,547023	0,04105413	0,2916039	0,498118
Błąd standard.	0,7411422	0,01520003	0,1808174	0,1742195
t(563)	-3,436618	2,700925	1,612699	2,85914
poziom p	0,0006324971	0,007122911	0,1073703	0,004405313
-95%CL	-4,002764	0,01119844	-0,06355507	0,1559184
+95%CL	-1,091281	0,07090981	0,646763	0,8403176
Chi-kwadrat Wald	11,81034	7,294995	2,600798	8,174683
poziom p	0,0005897734	0,006918336	0,1068198	0,004250655
Iloraz szans z jedn.	0,07831449	1,041909	1,338573	1,645621
-95%CL	0,01826508	1,011261	0,9384224	1,168731
+95%CL	0,335786	1,073484	1,90935	2,317103
Iloraz szans zakr.		2,272959	1,338573	1,645621
-95%CL		1,251032	0,9384224	1,168731
+95%CL		4,129665	1,90935	2,317103

Klasyfikacja przypadków (wcgs-min1)			
Il. szans: 2,0631 % poprawnych: 59,61%			
	Przew.	Przew.	Procent
Obserw.	0,000000	1,000000	Popraw.
0,000000	232	80	74,35897
1,000000	149	106	41,56863

Rys. 16. Wyniki analizy logistycznej przy użyciu predyktora wieku w skali interwałowej.

Porównując ze sobą wyniki trzech zbudowanych modeli logistycznych możemy stwierdzić, że jakkolwiek status palenia oraz wiek badanego mają istotny statystycznie wpływ na prawdopodobieństwo wystąpienia choroby wieńcowej i nie można negować istnienia zależności między tymi zmiennymi, to zastosowanie ich w modelu logistycznym nie jest wystarczające dla celów prognostycznych. Należy poszukiwać innych dodatkowych czynników, które pozwolą polepszyć zdolność predykcyjną. Użyta przez nas metoda to tzw. szybka regresja logistyczna dostępna z menu *Statystyka/Zaawansowane modele liniowe i nieliniowe/Estymacja nieliniowa*. Bardziej zaawansowane procedury dostępne są w menu *Statystyka/Zaawansowane modele liniowe i nieliniowe/Uogólnione modele liniowe i nieliniowe GLZ z logitową funkcją wiążącą*. Pozwalają one na dogłębną ocenę takich efektów między predyktorami, jak mediacja, interakcja i uwikłanie. Omówienie tych zagadnień wykracza poza ramy niniejszego opracowania, a zainteresowanego nimi Czytelnika odsyłam do pozycji literaturowych [1, 5].

Techniki data-mining

Omówione do tej pory metody statystyczne nie stanowią jedynego możliwego podejścia do analizy zmiennych jakościowych. W badaniach medycznych coraz częściej wykorzystywane są techniki oparte na sztucznej inteligencji. Dla przykładu pakiet *STATISTICA* pozwala na konstruowanie szerokiej gamy sztucznych sieci neuronowych, które mogą być

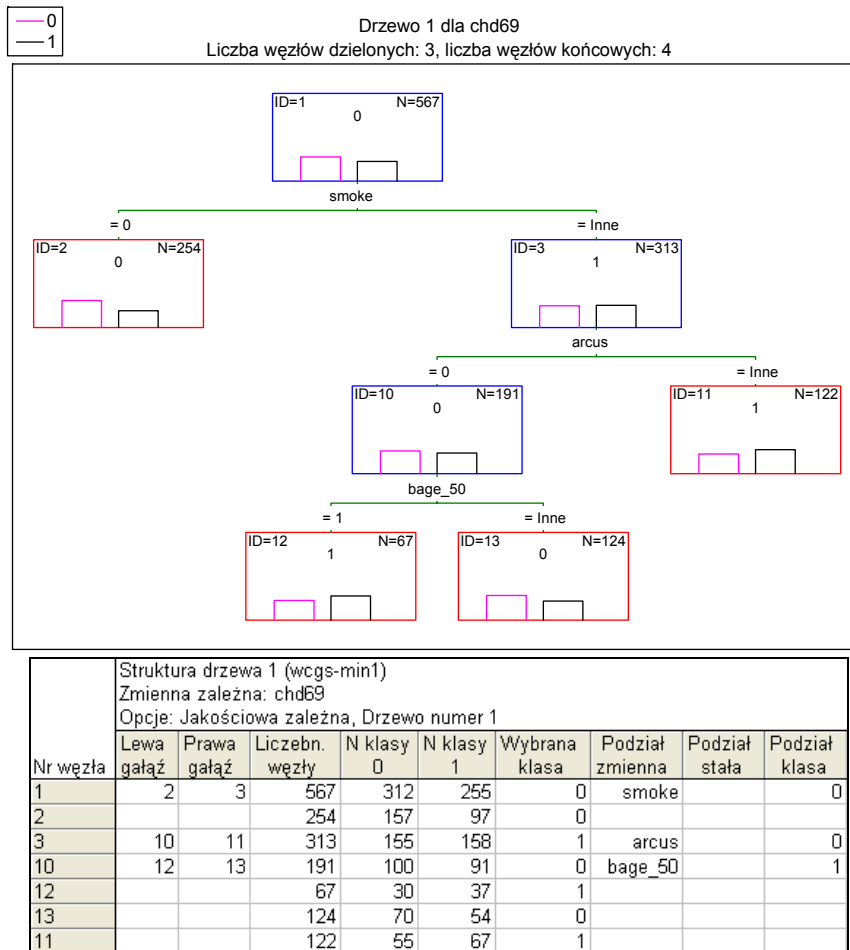


z powodzeniem wykorzystane do doboru istotnych dla modelowania predyktorów oraz prognozowania wartości zmiennych zależnych. Coraz bardziej znana staje się technika oparta na teorii zbiorów przybliżonych (*Rough Sets*), pozwalająca na proste generowanie reguł logicznych opisujących w dowolnych skalach pomiarowych zależności między zmiennymi.

Nasz krótki przegląd analizy zmiennych jakościowych zakończymy krótkim omówieniem najprostszych metod *data miningu* (coraz popularniejszy staje się spolszczony termin „zglębianie danych”). Zglębianie danych definiujemy jako proces automatycznego lub półautomatycznego badania dużych ilości danych w celu znalezienia istotnych zależności, wzorców i reguł. Należy podkreślić fakt, że wnioski uzyskiwane metodami zglębiania danych mają charakter indukcyjny. Oznacza to, że u źródła tworzonych modeli nie tkwią apriorycznie przyjęte abstrakcyjne teorie, lecz sama struktura analizowanych zbiorów danych. A więc w przeciwieństwie do technik statystycznych, gdzie weryfikowaliśmy hipotezy konstruowane a priori na podstawie założeń merytorycznych wynikających z naszej wiedzy o przedmiocie badań, metody zglębiania danych identyfikują systematyczne relacje („układy zależności”) występujące w zebranych wynikach pomiarów. Pośród szerokiej gamy rozmaitych technik data mining jedną z najbardziej przydatnych w badaniach medycznych wydaje się być technika drzew klasyfikacyjno-regresyjnych C&RT [7, 8]. Podstawową właściwością drzew jest ich hierarchiczna struktura. Oznacza to m.in., że wszystkie kolejne podziały są w pełni zależne od podziałów poprzednich, co wpływa na wysoką zmienność uzyskiwanych wyników. Zasadniczą zaletą drzew decyzyjnych jest natomiast brak jakichkolwiek założeń wstępnych dotyczących rozkładów danych. Spróbujemy zbadać, czy analizowane do tej pory metodami statystycznymi zmienne SMOKE, ARCUS, BAGE50 (lub AGE lub AGE) mogą służyć do chociażby częściowego prognozowania wystąpienia choroby wieńcowej. Do konstrukcji drzewa wybraliśmy technikę C&RT wyczerpującego poszukiwania podziałów jednowymiarowych dla predyktorów skategoryzowanych i porządkowych z opcjami jednakowego kosztu błędnej klasyfikacji, oceną jakości klasyfikacji przy użyciu miary Giniego, szacowanych prawdopodobieństw apriorycznych z regułą zatrzymania „przytnij przy błędzie złej klasyfikacji” [3, 4]. Zatrzymanie to polega na kontynuacji podziału drzewa do chwili, gdy wszystkie wierzchołki grafu będą „czyste” zgodnie z przyjętym kryterium dopuszczalnego poziomu „zanieczyszczenia”. W przypadku naszej bazy, zawierającej 569 badanych, przyjęcie dopuszczalnej frakcji błędnej klasyfikacji na poziomie 0,1 powoduje, że każdy wierzchołek terminalny zawierający co najwyżej 57 błędnie sklasyfikowanych obiektów będzie traktowany jako wierzchołek czysty. Można oczywiście polepszyć jakość klasyfikatora, zmniejszając parametr dopuszczalnej frakcji błędnej klasyfikacji do na przykład 0,01, lecz należy wtedy liczyć się ze wzrostem wielkości drzewa decyzyjnego. Co więcej, można oczywiście zażądać idealnie czystego podziału, jednakże uzyskamy wtedy niezwykle skomplikowaną strukturę drzewa, która aczkolwiek może doprowadzić do idealnej klasyfikacji zbioru uczącego, jednak nie będzie nadawała się do uogólnień. Podobnie można przyjąć wyższą wagę dla popełnianego błędu braku rozpoznania CHD z chwilą, gdy w rzeczywistości choroba występowała. Na rys. 17 przedstawiamy strukturę drzewa decyzyjnego dla podanych uprzednio parametrów. Jak widać, drzewo to zawiera cztery



wierzchołki terminalne „czyste” w sensie przyjętego kryterium. Pierwszy poziom podziału wykorzystuje zmienną SMOKE do podziału wierzchołka 1 (zawierającego 567 obiektów) na wierzchołki 2 i 3. Wierzchołek 2 zawiera 157 obiektów bez choroby wieńcowej, a tylko 97 z tą chorobą, i dlatego określony jest jako reprezentujący stan CHD=0. W przeciwieństwie do niego wierzchołek 3 jest powiązany ze stanem CHD=1 (155 przypadków z CHD=0, 158 z CHD=1).



Rys. 17. Struktura drzewa klasyfikacyjnego opartego na trzech predyktorach wyrażonych w skali nominalnej.

Przyjrzyjmy się teraz zbiorczej analizie utworzonego drzewa. Jej wyniki zamieszczamy w postaci macierzy klasyfikacyjnej zamieszczonej na rys. 18.

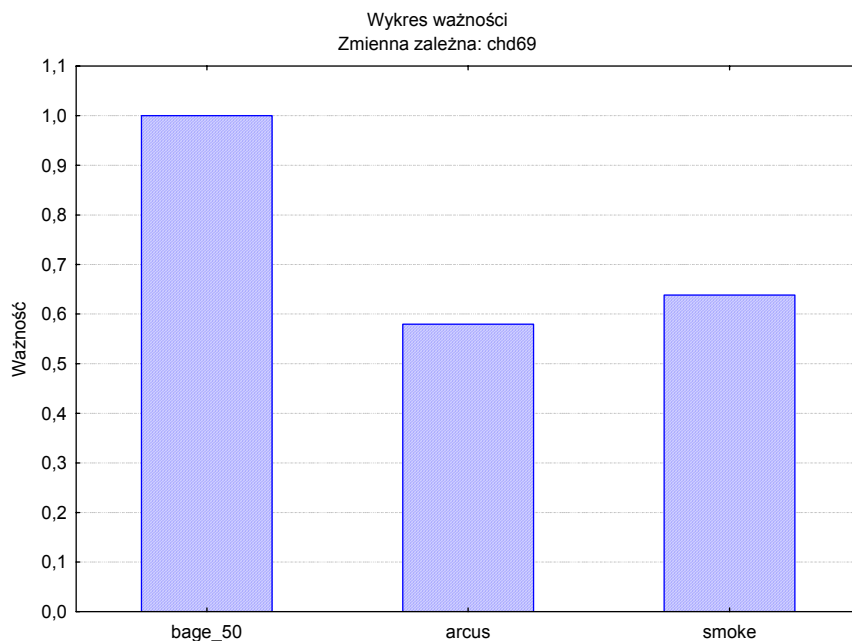
Jak łatwo zauważyć, wytworzone drzewo sklasyfikowało prawidłowo w klasie zerowej (brak choroby wieńcowej) 227 na 312 możliwych przypadków (72,8%), w klasie pierwszej odpowiednio 104 na 255 możliwych (40,8%). Ogólnie zatem prawidłowy poziom klasyfikacji realizowany przez drzewo wynosi około 58% przypadków.



Macierz klasyfikacji 1 (wcgs-min1)				
Zmienna zależna: chd69				
Opcje: Jakościowa zależna, Próba do analizy				
	Obserw.	Przewidywana 0	Przewidywana 1	Łącznie w wierszu
Numer	0	227	85	312
Procent z kolumny		60.05%	44.97%	
Procent z wiersza		72.76%	27.24%	
Procent z ogółu		40.04%	14.99%	55.03%
Numer	1	151	104	255
Procent z kolumny		39.95%	55.03%	
Procent z wiersza		59.22%	40.78%	
Procent z ogółu		26.63%	18.34%	44.97%
Liczba	Ogół grup	378	189	567
Procent łącznie		66.67%	33.33%	

Rys. 18. Macierz klasyfikacyjna dla danych uczących.

Jak widać, jakość klasyfikacyjna metody C&RT pokrywa się niemal idealnie z wynikami uzyskanymi przy użyciu metody logistycznej (rys. 14). Pośród analizowanych predyktorów największą rolę odgrywa wiek (BAGE50), najmniejszą zaś występowanie u badanego ARCUS SENILIS (rys. 19).



Rys. 19. Tablica stopnia ważności poszczególnych węzłów w decyzji o przynależności do danej klasy oraz histogram wag decyzyjnych poszczególnych predyktorów.

Musimy pamiętać, że w rzeczywistości, gdy będziemy dokonywali klasyfikacji przypadków testujących jakość rozpoznawania będzie niższa i dlatego możemy wyciągnąć z analizy wnioski, że niestety użyte przez nas zmienne opisujące są niewystarczające do zbudowania solidnego narzędzia klasyfikacyjnego. Niemniej jednak badane przez nas predyktory wykazują umiarkowany związek z występowaniem choroby wieńcowej. Zastępując zmienną BAGE50 zmiennymi AGE lub AGE, uzyskamy wnioski bardzo zbliżone do tych uzyskanych metodą analizy logistycznej. Warto zatem pogłębić analizę, zmieniając na przykład parametr kosztów błędnej klasyfikacji lub wykorzystując procedurę interakcyjnej budowy drzew C&RT.



Uwagi końcowe

Opisane w pracy techniki nie stanowią wyczerpującego przeglądu możliwych do użycia metod analizy danych jakościowych. Nie wspomnieliśmy tu o klasyfikatorach bayesowskich, analizie opartej na funkcjach sklepanych odcinkowo liniowych (*MARSpline*) lub wektory nośne (SVM – *Support Vector Machines*), analizie korespondencji, nie rozwinęliśmy tematu zbiorów przybliżonych (*Rough Sets*), sztucznych sieci neuronowych i wielu innych metod. Chcieliśmy pokazać jedynie najprostsze przykłady, w jaki sposób można uzyskać z danych pomierzonych w najniższej skali pomiarowej znacznie więcej informacji, niż oferują to powszechnie stosowane techniki tabelaryzacji jedno i wielodzielczej. Każda z technik bazuje na innych założeniach i dlatego uwypukla nieco inne właściwości zebranych danych. Przed wyciągnięciem ostatecznych wniosków korzystne jest zatem przyglądnięcie się danym różnymi metodami, co zapewni bardziej dokładną ich interpretację.

Literatura

1. Vittinghoff E., Glidden D.V., Shiboski S.C., McCulloch C.H.E. (2005). *Regression Methods In Biostatistics. Linear, Logistic, Survival and Repeated Measures Models*. Springer Science + Business Media, New York.
2. Mehta C.R., Patel N.R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association* 78(382):427-434.
3. Stanisław A. (2007). *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 3. Analizy wielowymiarowe*. Statsoft, Kraków.
4. Stanisław A. (2007). *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 2. Modele liniowe i nieliniowe*. Statsoft, Kraków.
5. Kleinbaum D.G., Klein M. (2002). *Logistic Regression. A self-learning text. Second edition*, Springer Science + Business Media, New York.
6. Hosmer D.W., Lemeshow S. (2000). *Applied Logistic Regression Second Edition*. Wiley, New York.
7. Loh W.-Y., Shih Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7, 815-840.
8. Loh W.-Y., Vanichestakul N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association*, 83, 715-728.