



CO MOŻNA WYCISNAĆ Z TYCH DANYCH?

*Andrzej Stanisław, Collegium Medicum Uniwersytetu Jagiellońskiego w Krakowie,
Zakład Biostatystyki i Informatyki Medycznej*

Lekarze i kliniki dysponują coraz większą liczbą danych. Jednocześnie rośnie problem poprawnego i jak najszybszego otrzymywania wyników analiz. W związku z tym coraz większego znaczenia nabiera dostęp do efektywnych narzędzi wspomaganie analizy danych. Głównym celem tego opracowania jest zaprezentowanie wybranych narzędzi wspomagających prowadzenie analizy danych nominalnych w programie *STATISTICA 8*. Analizowane dane zawierają informacje o: stosowaniu antykoncepcji, wieku, wykształceniu oraz pragnieniu posiadania dzieci w przyszłości. Ciąg analiz rozpoczyna test chi-kwadrat dla tabel wielodziedzielczych (kontyngencji). Kolejno do tych samych danych wykorzystywana jest analiza korespondencji, model regresji logistycznej oraz analiza log-liniowa. Bardziej szczegółowo opisane są różne analizy logitowe budowane z wykorzystaniem uogólnionego modelu liniowego. Referat nie omawia szczegółowo metod oraz założeń wykorzystywanych analiz, a jedynie podawana jest krótkka ich charakterystyka. Pokazane są głównie okna wynikowe i wynikające stąd powiązania między analizowanymi zmiennymi. Okazuje się, że nie wystarcza tradycyjne chi-kwadrat. Kolejne głębsze analizy odkrywają coraz bardziej interesujące powiązania między danymi. Tych zależności nie dały nam proste analizy tabel wielodziedzielczych.

Ogólna charakterystyka analizowanych danych

Prezentowane dane to fragment szerszego badania dotyczącego stosowaniem antykoncepcji. Rozpatrywane są, dla ponad 1000 osób, następujące cztery zmienne nominalne:

- ◆ **Antykoncepcja** – przyjmująca wartość: 1 – gdy kobieta stosuje antykoncepcję, a 0 w przeciwnym przypadku.
- ◆ **Dzieci** – przyjmująca wartość: 1 – jeżeli kobieta pragnie w przyszłości urodzić dzieci, a 0 w przeciwnym przypadku.
- ◆ **Wykształcenie** – określające dwie grupy (Wyższe, Niższe).
- ◆ **Wiek** – określający cztery grupy wiekowe (<25, 25–29, 30–39, 40–49).

Oryginalne dane wykorzystywane w opisywanym przykładzie zostały zapisane w arkuszu programu Excel. Fragment tych danych pokazany jest na rys. 1. W programie *STATISTICA 8* można taki arkusz otworzyć bezpośrednio, bez konieczności wcześniejszego importowania. Na otwartym w ten sposób arkuszu można następnie wykonywać analizy



i tworzyć wykresy, dokładnie tak samo jak w środowisku *STATISTICA*. Drugi sposób to wcześniejsze zaimportowanie danych z pliku źródłowego do formatu arkusza *STATISTICA*. W początkowej części analizowanego przykładu tak właśnie zrobiono.

Antykoncepcja	Dzieci		Antykoncepcja	Wiek		Antykoncepcja	Wykształcenie
Tak	Nie		Tak	<25		Tak	Wyższe
Nie	Tak		Nie	30-39		Nie	Niższe
Tak	Tak		Tak	40-49		Tak	Niższe
Tak	Nie		Tak	25-29		Tak	Wyższe
Tak	Nie		Tak	<25		Tak	Wyższe
Nie	Tak		Nie	40-49		Nie	Wyższe
Tak	Nie		Tak	30-39		Tak	Niższe
Nie	Nie		Nie	25-29		Nie	Niższe
Nie	Tak		Nie	<25		Nie	Niższe
Nie	Tak		Nie	<25		Nie	Niższe
Tak	Nie		Tak	25-29		Tak	Wyższe

Rys. 1. Fragment arkusza kalkulacyjnego zawierające surowe dane do analizy.

Dla bardziej zaawansowanych analiz zmodyfikowano jednak arkusz danych. Zamiast analizować oryginalne dane dokonano agregacji, grupując przypadki z identycznymi cechami. Wynikowy arkusz jest o wiele mniejszy, zawiera tylko 16 różnych grup (Wiek(4) × Wykształcenie(2) × Pragnienie dzieci(2)). Grupy te i ich liczebności pokazuje rys. 2.

Tabela liczebności (Liczebności)				
Liczebność oznacz. komórek > 10 (Nie oznaczono sum brzegowych)				
Wiek	Wykształcenie	Dzieci	Antykoncepcja Nie	Antykoncepcja Tak
40-49	Niższe	Nie	46	48
40-49	Niższe	Tak	35	6
40-49	Wyższe	Nie	12	31
40-49	Wyższe	Tak	8	8
30-39	Niższe	Nie	77	80
30-39	Niższe	Tak	112	33
30-39	Wyższe	Nie	68	78
30-39	Wyższe	Tak	118	46
25-29	Niższe	Nie	19	10
25-29	Niższe	Tak	60	14
25-29	Wyższe	Nie	65	27
25-29	Wyższe	Tak	155	54
< 25	Niższe	Nie	10	4
< 25	Niższe	Tak	53	6
< 25	Wyższe	Nie	50	10
< 25	Wyższe	Tak	212	52

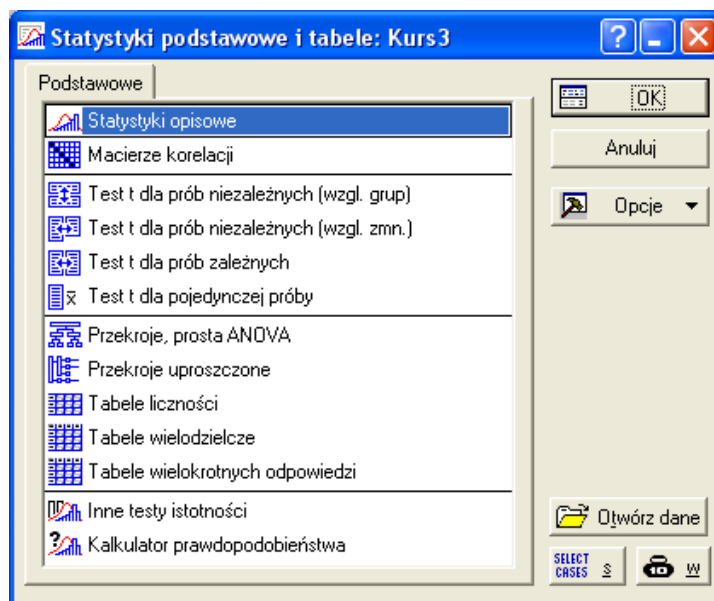
Rys. 2. Wynik agregacji oryginalnych danych.

Analiza testem chi-kwadrat

Ze sposobu zapisu danych wynika, że badaczowi chodziło o opisanie, jak stosowanie antykoncepcji jest zależne od wieku, wykształcenia oraz pragnienia posiadania dzieci



ankietowanych kobiet w przyszłości. Punktem wyjścia do takich analiz danych jest zestawienie ich w tabeli wielodzzielczej. Następnie analizujemy tak utworzone tabele za pomocą testu niezależności chi-kwadrat oraz wyliczamy dostępne tam statystyki (V Cramera, Φ -Yule'a itd.) informujące o sile związku pomiędzy zmiennymi jakościowymi. W pakiecie *STATISTICA 8.0* analizę tabel wielodzzielczych możemy przeprowadzić, wybierając w module *Statystyki podstawowe i tabele* opcję *Tabele wielodzzielcze*. Sytuacja ta pokazana jest na rys. 3.



Rys. 3. Okno wyboru analiz dotyczących tabel wielodzzielczych.

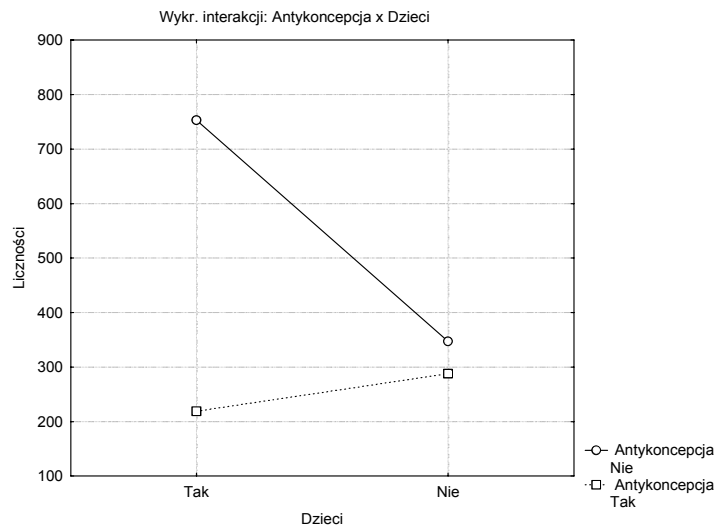
Wnioski z analizy testem chi-kwadrat

Wstępne analizy testem chi-kwadrat prezentowanych danych umożliwiają wstępną statystyczną i merytoryczną ocenę zebranych danych

1. Z pokazanego na rys. 4 arkusza wynikowego i wykresu możemy wnioskować, że występuje istotne powiązanie między pragnieniem posiadania w przyszłości dzieci a stosowaniem antykoncepcji ($p = 0,0000$ $F_i = 0,24$). Wśród pragnących w przyszłości mieć dzieci istotnie większy procent kobiet nie stosuje antykoncepcji. Dla niepragnących mieć dzieci ta różnica nie jest aż tak olbrzymia. Obliczony iloraz szans $OR = 2,87$ wskazuje, że szansa stosowania antykoncepcji wśród kobiet niepragnących mieć dzieci jest ponad 2,5 razy większa od szansy zastosowania antykoncepcji wśród kobiet pragnących w przyszłości mieć dzieci.



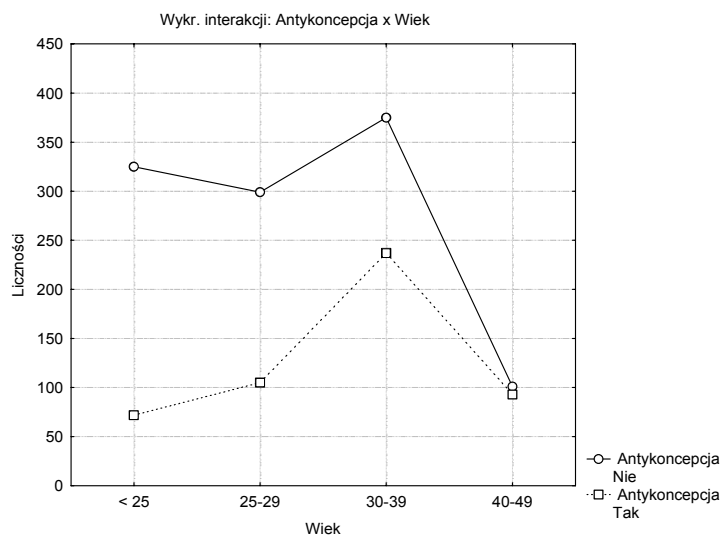
Statystyki:	Statystyki: Antykoncepcja(2,		
	Chi-kwadr.	df	p
Chi kwadrat Pearso	92,75263	df=1	p=0,0000
Chi ² NW	91,78096	df=1	p=0,0000
Chi ² Yatesa	91,69593	df=1	p=0,0000
dokt. Fishera, 1-stronny			-----
2-stronny			
Chi ² McNemara (A/D)	28,19893	df=1	p=,00000
(B/C)	205,8284	df=1	p=0,0000
Fi dla tabel 2 x 2	-,240770		
Korelacje tetrachoryczn	-,382252		
Wsp. kontyngencji	,2340811		



Rys. 4. Arkusz wyników i wykres testu chi-kwadrat dla *Antykoncepcji* i *Pragnienia posiadania dzieci*

2. Z pokazanego na rys. 5 arkusza wynikowego i wykresu możemy wnioskować, że występuje istotna zależność między strukturą wiekową a stosowaniem antykoncepcji ($p = 0,0000$ $F_i = 0,22$). Wydaje się, że liczba stosujących antykoncepcję wzrasta z wiekiem do okresu 30–39 lat, po którym następuje spadek. Wśród kobiet młodych (<25) zdecydowanie mniej stosuje antykoncepcję. W celu otrzymania bardziej przejrzystych wyników można porównywać procenty w poszczególnych grupach wiekowych oraz obliczyć na piechotę odpowiednie ilorazy szans.

Statystyki:	Statystyki: Antykoncepcja(2,		
	Chi-kwadr.	df	p
Chi kwadrat Pearso	77,89126	df=3	p=,00000
Chi ² NW	79,62600	df=3	p=,00000
Fi	,2206401		
Wsp. kontyngencji	,2154579		
V Craméra	,2206401		



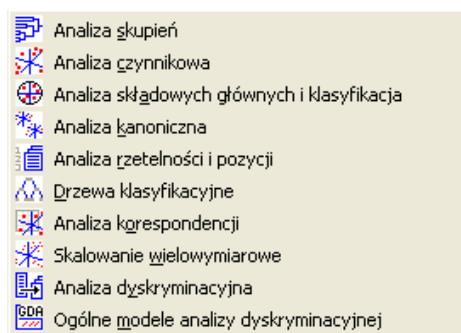
Rys. 5. Arkusz wyników i wykres testu chi-kwadrat dla *Antykoncepcji* i *Wiek*.

3. Nie występuje natomiast istotne powiązanie między wykształceniem a stosowaniem antykoncepcji ($p = 0,3586$). Wskazywałoby to na możliwość pominięcia tej zmiennej w dalszych analizach. Ale czy słusznie?



Zastosowanie analizy korespondencji

Omawiane powyżej statystyki informują nas tylko o istotności i sile związku pomiędzy zmiennymi jakościowymi, nie opisują jednak charakteru powiązań pomiędzy kategoriami analizowanych zmiennych jakościowych. Analiza korespondencji to opisowa i eksploracyjna technika, która dostarcza nam informacji o strukturze powiązań między kolumnami i wierszami tabeli wielodzzielczej. W pakiecie *STATISTICA 8.0* analizę korespondencji możemy przeprowadzić, wybierając w menu *Statystyka* ciąg opcji *Wielowymiarowe techniki eksploracyjne\Analiza korespondencji*. Sytuacja ta pokazana jest na rys. 6.



Rys. 6. Menu wyboru *Analizy korespondencji*.

Analiza korespondencji dostarcza informacji podobnych w interpretacji do wyników analizy czynnikowej, dotyczących jednak zmiennych jakościowych. Analiza statystyk i wykresów – zaproponowanych przez tę metodę – umożliwi na proste i intuicyjne wnioskowanie o powiązaniach zachodzących pomiędzy kategoriami zmiennych.

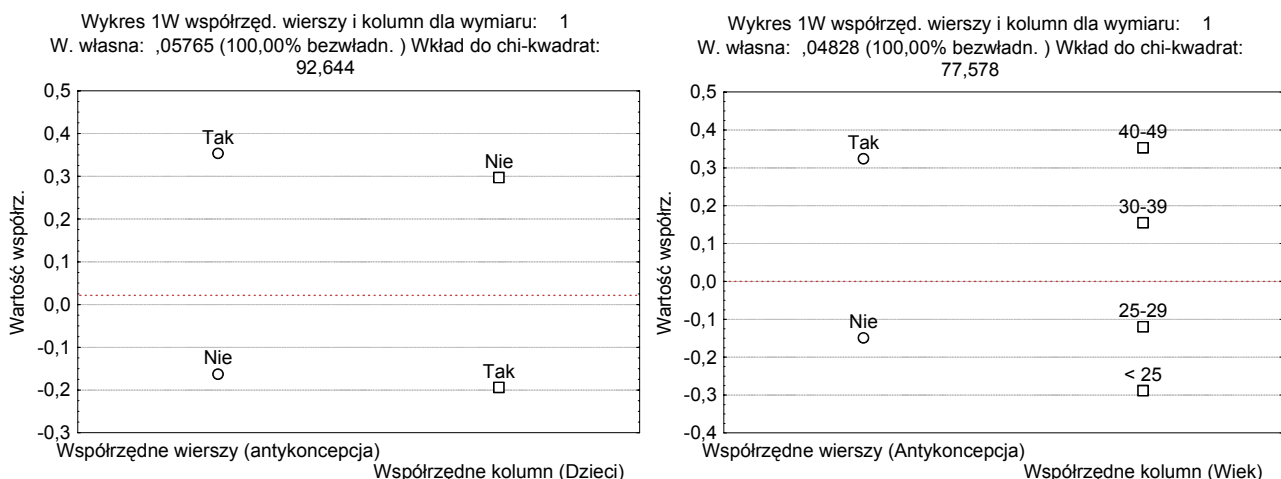
Procedura analizy korespondencji przebiega w 7 etapach. Większość etapów przeprowadzamy najpierw dla kategorii jednej zmiennej (wiersze), potem dla kategorii drugiej zmiennej (kolumny). Najważniejsze kroki to:

1. Wyznaczenie profili wierszowych i kolumnowych.
2. Wyznaczenie masy wiersza i kolumny.
3. Obliczenie odległości między wierszami (kolumnami) za pomocą metryki chi-kwadrat. Aby przeprowadzić analizę profili, wylicza się odległość między nimi za pomocą ważonej metryki euklidesowej nazywanej metryką chi-kwadrat. Z metryką chi-kwadrat blisko związane jest również pojęcie bezwładności. Termin bezwładność jest używany w analizie korespondencji analogicznie do występującego w statystyce pojęcia wariancji. Całkowita bezwładność jest bowiem miarą rozproszenia profili wokół odpowiednich przeciętnych profili. Można pokazać, że bezwładność dla wierszy jest równa bezwładności dla kolumn. Jeśli bezwładność jest bliska zeru, wtedy różnica między profilami a profilem przeciętnym jest niewielka, co oznacza niewielkie rozproszenie wokół profilu przeciętnego. Z kolei duża wartość bezwładności oznacza duże rozproszenie wokół profilu przeciętnego. Z powiązania bezwładności z wartością testu chi-kwadrat wynika, że im mniejsza bezwładność, tym mniejsza szansa wystąpienia istotnego powiązania między wierszami i kolumnami tabeli wielodzzielczej.



4. Przedstawienie profili wierszowych (kolumnowych) w przestrzeni generowanej przez kolumny (wiersze) macierzy korespondencji.
5. Wyznaczenie przeciętnych profili wierszowych i kolumnowych.
6. Redukcja wymiaru przestrzeni, która najlepiej odpowiada analizowanym profilom. Najczęściej jest to dwu- lub trójwymiarowa przestrzeń. Następnie dokonujemy rotacji tak utworzonego układu, aby maksymalizować wariację wyjaśnioną przez kolejne współrzędne tej przestrzeni, W ten sposób definiujemy układ współrzędnych, w który rzutowane będą punkty odpowiadające kolejnym wierszom i kolumnom. W praktyce rozpatrujemy dwa lub trzy wektory osobliwe, zarówno dla kolumn, jak i wierszy. Możemy wówczas informację o podobieństwie między wierszami (kolumnami) przedstawić na prostym dwu- lub trójwymiarowym rysunku.
7. Utworzenie wspólnego wykresu profili wierszowych i kolumnowych za pomocą współrzędnych głównych. Gdy zdecydujemy się na wyodrębnienie odpowiedniej liczby wymiarów, możemy obliczyć współrzędne profili wierszowych i kolumnowych w nowym układzie współrzędnych (tzw. współrzędnych głównych). Umożliwia to utworzenie wykresu obrazującego położenie punktów reprezentujących wiersze i kolumny. Ten wspólny wykres można wykorzystać dla znalezienia grup (nieokreślonych a priori) ilustrujących zależności między wierszami i kolumnami.

Dla wstępnej orientacji co do rodzaju i kierunku ewentualnego powiązania cech zostały utworzone dwa wykresy profili wierszowych i kolumnowych. Na rys. 7 mamy przedstawione w nowym układzie współrzędnych profile wierszy i kolumn dwóch analizowanych cech. Rysunek po lewej stronie pokazuje położenie punktów reprezentujących dwie cechy: *Antykoncepcja* i *Pragnienie posiadania w przyszłości dzieci*, a rysunek po prawej stronie położenie punktów reprezentujących *Antykoncepcję* i *Wiek*.



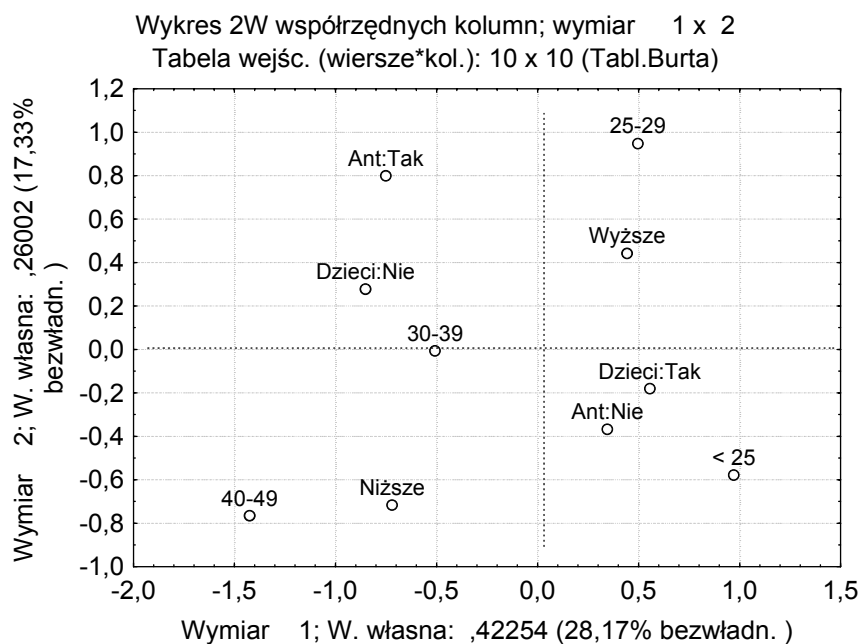
Rys. 7. Wykresy 1W dla poszczególnych osi.

Zamieszczony powyżej wykres pozwala zaobserwować, że pierwsza oś (przerwana linia), mająca największy udział w bezwładności, wyróżnia na obu rysunkach dwie grupy. Niektóre podziały są oczywiste (stosujące i niestosujące antykoncepcji; pragnące i niepragnące mieć dzieci). Grupy związane z wiekiem potwierdzają spostrzeżenia poczynione



w analizie za pomocą testu chi-kwadrat. Relatywnie więcej stosujących antykoncepcję jest wśród starszych kobiet (30–39 oraz 40–49). Natomiast więcej niestosujących występuje w młodszych grupach wiekowych (<25 oraz 25–30).

Jak widzimy, procedura ta może wydawać się mało przydatna dla małych tablic, jednakże jest szczególnie przydatna dla dużych tablic, ułatwiając ich prezentację i interpretację. Bardziej interesujące wyniki otrzymamy, badając związki między wszystkim analizowanymi zmiennymi jakościowymi. Umożliwia nam to wielowymiarowa analiza korespondencji. Jest ona naturalnym rozszerzeniem prostej analizy korespondencji na zagadnienia o liczbie zmiennych większej od dwóch. Przykładowy wykres współrzędnych kolumn pokazuje rys. 8.



Rys. 8. Wykresy współrzędnych kolumn dla sytuacji wielowymiarowej.

Jak widzimy, pierwsza oś (pionowa) rozdziela grupę kobiet, które stosują antykoncepcję (po lewej stronie), od pacjentek, które jej nie stosują. Pragnienie dziecka w przyszłości zdecydowanie zmniejsza szansę stosowania antykoncepcji. Punkt reprezentujący *Dzieci:Tak* leży bardzo blisko punktu *Antykoncepcja:Nie*. Natomiast punkt *Dzieci:Nie* leży bliżej punktu *Antykoncepcja:Tak*. Punkty opisujące młodsze grupy wiekowe leżą po tej samej stronie (prawa), co punkt reprezentujący kobiety niestosujące antykoncepcji. Zauważamy jednak, że punkt reprezentujący grupę wiekową 25–29 lat jest bardziej oddalony od punktu *Antykoncepcja:Nie* niż punkt reprezentujący grupę najmłodszą. Jest to prawdopodobnie związane z wykształceniem, który wydaje się być czynnikiem modyfikującym zauważone efekty.

Druga oś (pozioma) związana jest głównie z podziałem ze względu na wykształcenie. Wykształcenie wyższe wydaje się być związane z kobietami w wieku od 25 do 30 lat. Punkt reprezentujący tę grupę wiekową leży blisko punktu *Wyższe*. Natomiast wykształcenie niższe związane jest z najstarszą grupą wiekową (40–49) i najmłodszą, która



prawdopodobnie ze względu na swój wiek jest dopiero w okresie zdobywania wykształcenia. Pośrednia grupa wiekowa (30–39) leży w połowie na osi poziomej.

Wnioski z analizy korespondencji

- ◆ Potwierdzają wnioski z analizy testem chi-kwadrat.
- ◆ Wskazują na wystąpienie interakcji między analizowanymi zmiennymi.
- ◆ Wykształcenie, choć bezpośrednio niepowiązane ze stosowaniem antykoncepcji, wydaje się wpływać na sposób powiązania pozostałych zmiennych z antykoncepcją.

Model regresji logistycznej

Chęć lepszego poznania badanego zjawiska, czyli opisu charakteru i siły powiązania pomiędzy zmiennymi interesującymi badacza, sugeruje budowanie modeli podobnych do modeli regresji wielokrotnych. Takie modele, pod nazwą liniowe modele prawdopodobieństwa (LMP), budowano i analizowano jeszcze do niedawna (Amemiya 1977). W naszej sytuacji przyjmuje on postać:

$$\pi_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni}$$

gdzie: n_i – liczba obserwacji w i -tej grupie,

y_i – liczba kobiet stosujących antykoncepcję,

π_i – prawdopodobieństwo stosowania antykoncepcji,

X_j – zmienne, które podejrzewamy, że mają wpływ na prawdopodobieństwo stosowania antykoncepcji.

Oceny parametrów LMP, mające interpretację typową dla liniowego modelu regresji, otrzymujemy za pomocą klasycznej metody najmniejszych kwadratów lub uogólnionej metody najmniejszych kwadratów.

Patrzmy na y_i jak na realizację zmiennej losowej Y_i , która przyjmuje wartości 0, 1, ..., n_i . Rozkład Y_i jest rozkładem dwumianowym z parametrami n_i oraz π_i . Wartość oczekiwana i wariancja zmiennej są odpowiednio równe: $E(Y_i) = n_i \pi_i$ i $V(Y_i) = n_i \pi_i (1 - \pi_i)$. Wynika stąd, że LMP ma dwie zasadnicze wady:

- ◆ Klasyczna regresja może przewidzieć wartości Y , który są ujemne albo większe niż 1, co może dać nonsensowne predykcje Y .
- ◆ Reszty nie spełniają założenia o jednorodności wariancji (homoskedastyczności).

Zatem proponowany powyżej model jest niewłaściwy i nie będzie dalej omawiany.

Doskonałym modelem do analizowania omawianych zagadnień jest model regresji logistycznej. Ogólnie mówiąc, model regresji logistycznej jest formułą matematyczną, której możemy użyć do opisanego wpływu kilku zmiennych X_1, X_2, \dots, X_k na dychotomiczną zmienną Y , przyjmującą dwie wartości (sukces/porażka):

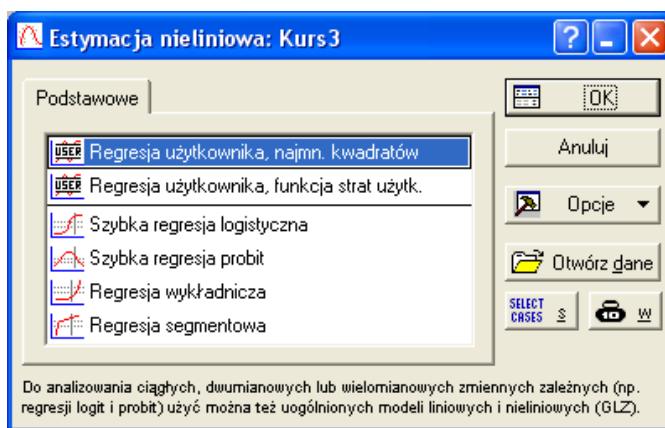


$$P(Y = 1 | x_1, x_2, \dots, x_k) = \frac{e^{\left(\alpha_0 + \sum_{i=1}^k \alpha_i x_i\right)}}{1 + e^{\left(\alpha_0 + \sum_{i=1}^k \alpha_i x_i\right)}}$$

gdzie: α_i to współczynniki regresji dla $i = 0, \dots, k$,
 x_i – zmienne niezależne (mieralne lub jakościowe) dla $i = 1, 2, \dots, k$.

Oceny współczynników otrzymujemy za pomocą metody największej wiarygodności. Oceny istotności poszczególnych zmiennych przeprowadzamy za pomocą testu t lub statystyki Walda. Oceny dopasowania modelu do danych przeprowadzamy za pomocą statystyki LR. Obliczana jest statystyka $-2\log$ (maksimum wiarygodności dla naszego modelu) i modelu tylko z wyrazem wolnym. Różnica dla dużych prób ma rozkład zbliżony do chi-kwadrat.

W pakiecie *STATISTICA 8.0* analizę regresji logistycznej możemy przeprowadzić, wybierając w menu *Statystyka* ciąg opcji *Zaawansowane modele liniowe i nieliniowe/Estymacja nieliniowa*. Sytuacja ta pokazana jest na rys. 9.



Rys. 9. Okno wyboru modułu *Regresja logistyczna*.

Dla naszych przykładowych danych otrzymujemy następujący arkusz wyników.

		Model: Regr. logistyczna (logit) N zer: 1100 jedynek: 500 (Liczności) Zmn. zal.: Antykoncepcja Strata: Największe prawd. bł.średnkw.skal. -2*log(wiarygodn.) dla tego mod=1867,838 wyraz wolny =2003,694 Całkowita strata: 933,91919506 Chi2(5)=135,86 p=0,0000					
N=1600		Stała B0	Dzieci	Wykształcenie	Wiek1	Wiek2	Wiek3
Ocena		-1,133	-0,833	0,325	0,389	0,909	1,189
Błąd standard.		0,187	0,117	0,124	0,176	0,165	0,214
t(1601)		-6,057	-7,091	2,621	2,214	5,520	5,546
poziom p		0,000	0,000	0,009	0,027	0,000	0,000
-95%CL		-1,500	-1,063	0,082	0,044	0,586	0,769
+95%CL		-0,766	-0,603	0,568	0,734	1,232	1,610
Chi-kwadrat Walda		36,693	50,280	6,868	4,901	30,465	30,759
poziom p		0,000	0,000	0,009	0,027	0,000	0,000
Iloraz szans z.jedn.		0,322	0,435	1,384	1,476	2,481	3,285
-95%CL		0,223	0,345	1,085	1,045	1,796	2,157
+95%CL		0,465	0,547	1,765	2,084	3,426	5,002

Rys. 10. Arkusz wyników dla modelu logistycznego.



Zawarte w tabeli wyniki modelowania umożliwiają statystyczną i merytoryczną ocenę zbudowanego modelu. Okazuje się, że otrzymaliśmy model bardzo dobrze dopasowany do danych (wartość p dla statystyki LR jest równa 0,0000) o następującej postaci:

$$P(Y = 1) = \frac{e^{(-1,133 - 0,833 \cdot Dzieci + 0,325 \cdot Wyk + 0,389 \cdot Wiek1 + 0,909 \cdot Wiek2 + 1,189 \cdot Wiek3)}}{1 + e^{(-1,133 - 0,833 \cdot Dzieci + 0,325 \cdot Wyk + 0,389 \cdot Wiek1 + 0,909 \cdot Wiek2 + 1,189 \cdot Wiek3)}}$$

Wnioski z analizy logistycznej

Wszystkie parametry strukturalne modelu istotnie różnią się od zera. Obliczone z próby oceny parametrów informują, że:

1. Prawdopodobieństwo stosowania antykoncepcji spada, jeśli kobieta pragnie w przyszłości urodzić dzieci. Szansa zastosowania antykoncepcji u kobiet niechcących w przyszłości mieć dzieci jest ponad 2 razy większa ($OR = 1/0,435 = 2,299$) niż szansa zastosowania antykoncepcji u kobiet pragnących mieć dzieci.
2. Wyższe wykształcenie jest czynnikiem stymulującym stosowanie antykoncepcji. Szansa zastosowania antykoncepcji u kobiet z wyższym wykształceniem jest około 1,4 razy większa ($OR = 1,384$) niż szansa zastosowania antykoncepcji u kobiet z wykształceniem niższym.
3. Zmienną *Wiek* (określającą cztery grupy wiekowe <25, 25–29, 30–39, 40–49) wprowadzono do modelu, tworząc trzy sztuczne zmienne *Wiek1* (dla 25–29), *Wiek2* (dla 30–39) oraz *Wiek3* (dla 40–49) reprezentujące odpowiednie grupy wiekowe. Jako poziom odniesienia przyjęto grupę najmłodszą (< 25). Okazuje się wówczas, że:
 - ◆ Szansa zastosowania antykoncepcji przez kobiety w wieku 25–29 lat jest około 1,5 razy większa ($OR = 1,476$) niż szansa zastosowania antykoncepcji przez kobiety w wieku <25 lat.
 - ◆ Szansa zastosowania antykoncepcji przez kobiety w wieku 30–39 lat jest około 2,5 razy większa ($OR = 2,481$) niż szansa zastosowania antykoncepcji przez kobiety w wieku <25 lat.
 - ◆ Szansa zastosowania antykoncepcji przez kobiety w wieku 40–49 lat jest około 3,3 razy większa ($OR = 3,285$) niż szansa zastosowania antykoncepcji przez kobiety w wieku <25 lat.

Wykorzystanie uogólnionego modelu liniowego

Jak już wspomniano, jedną z trudności zastosowania klasycznego modelu regresji wielokrotnej do modelowania prawdopodobieństwa był zakres przyjmowanych wartości. Prawdopodobieństwo przyjmuje wartości z przedziału od 0 do 1, natomiast wartości wyliczone dla dowolnego modelu regresji mogą przyjmować dowolną wartość rzeczywistą. Pierwszym krokiem do przezwyciężenia tych trudności było rozważanie szans zamiast prawdopodobieństwa. „Szansa” jest to stosunek prawdopodobieństwa, że jakieś zdarzenie wystąpi,



do prawdopodobieństwa, że to zdarzenie nie pojawi się. Dla danego przypadku A powyższą definicję możemy zapisać następującym wzorem:

$$S(A) = \frac{p(A)}{p(\text{nie}A)} = \frac{p(A)}{1 - P(A)}$$

Przykładowo, jeśli $p(A) = 0,8$, to $S(A) = 0,8/(1-0,8) = 4$. Oznacza to, że prawdopodobieństwo pojawienia się przypadku A jest 4 razy większe niż prawdopodobieństwa niepojawienia się tego przypadku. Mówimy też, że szansa wystąpienia przypadku A jest 4 do 1.

Szansa przyjmuje wartości od 0 do nieskończoności. Jej logarytm z kolei przyjmuje dowolną wartość rzeczywistą. Zatem zamiast modelować wartości prawdopodobieństw możemy modelować logarytm szans. Prowadzi to nas do zastosowania transformacji logitowej postaci:

$$\text{Logit } P = \ln \frac{P}{1-P} = \ln \frac{p(Y=1)}{1-P(Y=1)} = a_0 + \sum_{i=1}^k a_i x_i$$

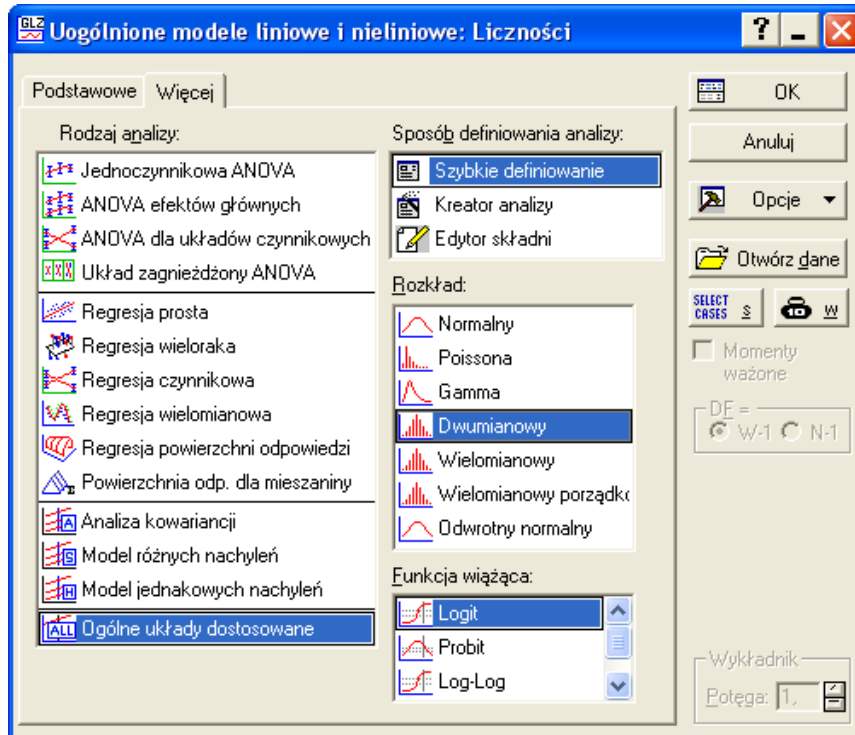
Ponieważ logit przyjmuje dowolne wartości rzeczywiste, możemy szukać powiązania ze zmiennymi niezależnymi w postaci liniowej pokazanej po prawej stronie powyższej równości.

Rozwój metod statystycznych w ostatnich dwudziestu latach pozwolił uogólnić podejście regresji liniowej na różne typy zmiennej zależnej. A wszystko to za sprawą uogólnionego modelu liniowego. Wykorzystując różne funkcje wiążące, uogólniony model liniowy rozszerza idee leżące u podstaw analizy regresji liniowej na następujące przypadki:

- ◆ Zmienna zależna może być dychotomiczna lub przyjmować skończoną liczbę wartości.
- ◆ Powiązanie między zmiennymi niezależnymi a zmienną zależną nie musi przyjmować postaci liniowej.

Dalszą analizę kontynuujemy z zastosowaniem ogólnego modelu liniowego. W programie *STATISTICA* służy do tego moduł *Uogólnione modele liniowe i nieliniowe* dostępny po wybraniu z menu *Statystyka* opcji *Zaawansowane modele liniowe i nieliniowe/Uogólnione modele liniowe i nieliniowe*. Sytuacja ta pokazana jest na rys. 11.

Przy przeprowadzaniu różnorodnych analiz korzystających z wielu modułów mogą zdarzyć się sytuacje, w których badacz chciałby przerwać analizę i „zamrozić” jej stan, tak aby mógł w dowolnym momencie powrócić do przerwanej analizy w tym samym miejscu. W nowej wersji programu *STATISTICA* dodano bardzo użyteczną możliwość zapisywania stanu przeprowadzanych analiz w pliku projektu analizy. Do projektu mogą zostać ponadto dołączone wykresy, skoroszyty, makra, raporty, rozpoczęte analizy oraz wyniki analiz. Po ponownym otworzeniu w programie zapisanego projektu można kontynuować analizy w tym samym miejscu, w którym zostały przerwane.



Rys. 11. Okno wyboru postaci uogólnionego modelu liniowego.

Jak widzimy, możemy dla logitów analizować modele podobne do modeli:

- ◆ jednoczynnikowej analizy wariancji postaci: $\text{Logit}(\pi_i) = \lambda + \alpha_i$,
- ◆ dwuczynnikowej analizy wariancji postaci: $\text{Logit}(\pi_{ij}) = \lambda + \alpha_i + \beta_j + (\beta\alpha)_{ij}$,
- ◆ regresji prostej liniowej postaci: $\text{Logit}(\pi_i) = \alpha + \beta x_i$,
- ◆ regresji wielokrotnej postaci: $\text{Logit}(\pi_i) = \alpha_i + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$,
- ◆ analizy kowariancji postaci: $\text{Logit}(\pi_i) = \alpha_i + \beta x_i$,
- ◆ różnych nachyleń postaci: $\text{Logit}(\pi_i) = \alpha_i + \beta_i x_i$.

Wynika stąd, że uogólniony model liniowy, z rozkładem dwumianowym i funkcją logit jako funkcją wiążącą, umożliwia badanie interakcji i bardziej zaawansowanych modeli dla naszych przykładowych danych. Oceny współczynników otrzymujemy za pomocą metody największej wiarygodności. Oceny istotności poszczególnych zmiennych przeprowadzamy za pomocą statystyki Walda. Oceny dopasowania modelu do danych przeprowadzamy za pomocą statystyki odchylenia D. Jest to miara niezgodności między obserwowanymi i dopasowanymi wartościami. Doskonale dopasowanie to $D = 0$. Dla dużych prób D jest aproksymowane rozkładem chi-kwadrat z $n - p$ stopniami swobody, gdzie n – liczba grup, a p – liczba parametrów. Oznaczmy przez $D(X_1/p_1)$ wartość odchylenia i stopnie swobody dla modelu zawierającego zmienną tylko X_1 oraz $D(X_1 + X_2/p_1+p_2)$ wartość odchylenia i stopnie swobody dla modelu zawierającego zmienne X_1 i X_2 . Mamy wówczas $\chi^2 = D(X_1/p_1) - D(X_1 + X_2/p_1+p_2)$ z liczbą stopni swobody $df = p_2$. Oznacza to, że statystyka D umożliwia też porównanie zagnieżdżonych modeli. Pamiętajmy jednak, że statystykę tę



stosujemy tylko dla danych zgrupowanych. Alternatywą może być statystyka χ^2 Pearsona, ale bez porównywania modeli.

Zainspirowani poprzednimi wynikami przeprowadzimy dwuczynnikową analizę uwzględniającą: *Wiek*, *Pragnienie posiadania dzieci w przyszłości* i ich interakcję. Poniższy arkusz wyników (rys. 12) dotyczy dwuczynnikowego modelu $\lambda + \alpha_i + \beta_j + (\beta\alpha)_{ij}$.

gdzie: λ – logit prawdopodobieństwa stosowania antykoncepcji dla kobiet poniżej 25 lat i które chcą mieć dzieci w przyszłości (poziom odniesienia),

α_i – efekt netto grup wiekowych w porównaniu do grupy <25 w tej samej kategorii pragnienia posiadania dzieci,

β_j – efekt netto „niechcących” mieć dzieci w stosunku do tych, które „chcą”.

Antykoncepcja - Oceny parametrów (Liczności)							
Rozkład: DWUMIANOWY							
F. wiążąca: LOGIT							
Efekt	Poziom Efekt	Kolumna	Ocena	Standard Błąd	Wald Stat.	p	NowaZm =exp(v3)
W.wolny		1	-1,51929	0,144965	109,8376	0,000000	0,218868
Dzieci	Tak	3	0,00000				1
Wiek	40-49	4	0,39714	0,340146	1,3632	0,242981	1,48757
Wiek	30-39	5	0,45066	0,194990	5,3415	0,020823	1,56934
Wiek	25-29	6	0,36816	0,200928	3,3573	0,066909	1,445068
Wiek	< 25	7	0,00000				1
Dzieci	Nie	2					
Wiek	40-49	8	1,36715	0,483419	7,9980	0,004683	3,924143
Wiek	30-39	9	1,09049	0,373285	8,5342	0,003485	2,975741
Wiek	25-29	10	0,26723	0,409144	0,4266	0,513661	1,306343
Wiek	<25	11	0,06400	0,330318	0,0375	0,846371	1,066092

Rys. 12. Arkusz wyników dla analizy dwuczynnikowej (*Wiek* i *Pragnienie posiadania dzieci*).

Dodatkowo wyliczono ilorazy szans. Na podstawie uzyskanych wyników możemy stwierdzić, że (zauważone wcześniej) zmiany szansy stosowania antykoncepcji wraz ze wzrostem wieku zależą od pragnienia posiadania dzieci w przyszłości. Dla kobiet pragnących mieć dzieci w przyszłości szanse w poszczególnych grupach różnią się minimalnie. Natomiast dla kobiet niepragnących mieć dzieci w przyszłości iloraz szans rośnie drastycznie wraz z wiekiem, osiągając w najstarszej grupie wartość OR = 3,92. Oznacza to, że szansa stosowania antykoncepcji wśród kobiet w wieku 40–49 jest prawie czterokrotnie większa od szansy stosowania antykoncepcji wśród kobiet poniżej 25 roku życia. Jak widzimy, wprowadzenie interakcji dało nowe spojrzenie na omawiane zagadnienie. Nie jest to jednak najlepszy model, nie uwzględnia bowiem efektu *Wykształcenia*.

W analizie przykładowych danych zastosowano wiele modeli uwzględniających analizy jedno- i wieloczynnikowe oraz modele analizy kowariancji. Nie sposób przedstawić szczegółowo wyników wszystkich analiz. Dla przykładu omówiliśmy powyżej analizę dwuczynnikową. Wyniki pozostałych analiz przedstawimy w zbiorczych tabelach. W tabelach tych istotne odchylenia zaznaczono czcionką pogrubioną.



Modele regresyjne

Poniższa tabela zawiera charakterystyki modeli regresyjnych:

Tabela 1. Zestawienie modeli „regresyjnych”.

Model	Logit(π)	Odchylenie	Stopnie swobody
Jedna linia	$\alpha + \beta x_i$	68,9	6
Linie równoległe	$\alpha_j + \beta x_i$	18,9	5
Dwie linie	$\alpha_j + \beta_j x_i$	9,14	4

Ponieważ można zauważyć, że logity wydają się wzrastać w przybliżeniu o tę samą wartość, jeśli przechodzimy z jednej grupy do następnej, podjęliśmy więc próbę weryfikowania modeli regresyjnych, wprowadzając nową zmienną ciągłą związaną z wiekiem. Przyjmuje ona wartości środkowe przedziałów klasowych (20; 27,5; 35 oraz 45). Wówczas z modelu regresji wynika, że logit prawdopodobieństwa stosowania antykoncepcji powiększa się o 0,061 dla każdego wzrostu *Wiek* o rok. Oznacza to, że szansa stosowania antykoncepcji powiększa się o 6,3% ($\exp(0,061) = 1,063$) dla każdego wzrostu o rok. W modelach tych uwzględniono również zmienną *Pragnienie posiadania dzieci* jako zmienną jakościową. Zawarte w tabeli wyniki pokazują, że najlepiej dopasowany do danych (najmniejsze nieistotne odchylenie) jest model dwóch linii o różnych nachyleniach.

Modele wieloczynnikowe

Poniższa tabela zawiera charakterystyki rozważanych modeli wieloczynnikowych. Dla zwięzłości i oszczędności zapisu przyjęto następujące oznaczenia:

- ◆ Zmienna Wiek – litera L i symbol α w modelu.
- ◆ Zmienna Wykształcenie – litera W i symbol β w modelu.
- ◆ Zmienna Dzieci – litera D i symbol γ w modelu.

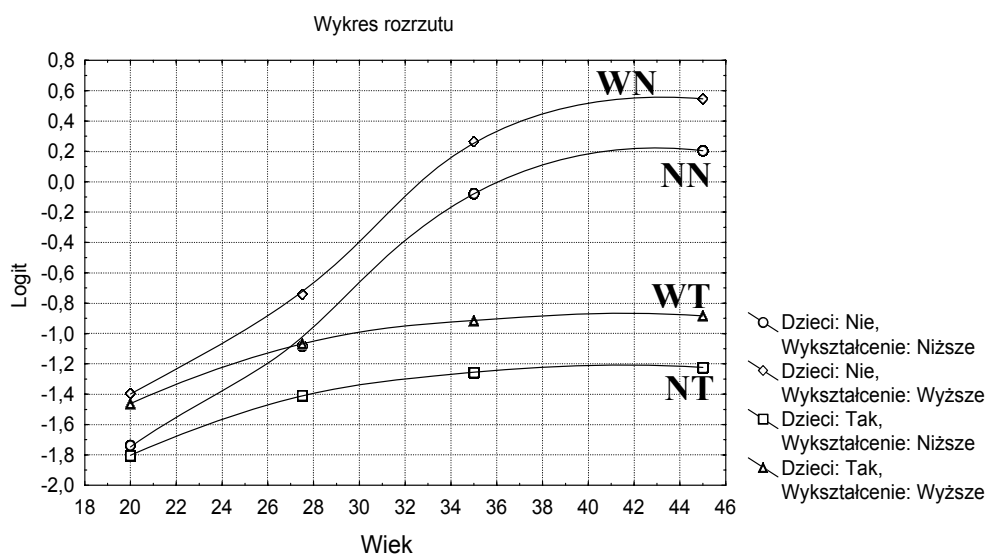
Tabela 2. Wybrane modele wieloczynnikowe.

Model	Logit(π)	Odchylenie	Stopnie swobody
Wiek (L)	$\lambda + \alpha$	86,6	12
Wykształcenie (W)	$\lambda + \beta$	165,1	14
Dzieci (D)	$\lambda + \gamma$	74,1	14
L + W	$\lambda + \alpha + \beta$	80,42	11
L + D	$\lambda + \alpha + \gamma$	36,9	11
W + D	$\lambda + \beta + \gamma$	73,9	13



Model	Logit(π)	Odchylenie	Stopnie swobody
$L \times W + L + W$	$\lambda + \alpha + \beta + (\beta\alpha)$	73	8
$L \times D + L + D$	$\lambda + \alpha + \gamma + (\gamma\alpha)$	20,1	8
$W \times D + W + D$	$\lambda + \beta + \gamma + (\gamma\beta)$	67,6	12
$L + W + D$	$\lambda + \alpha + \beta + \gamma$	29,9	10
$L \times W + L + W + D$	$\lambda + \alpha + \beta + \gamma + (\beta\alpha)$	23,2	7
$L \times D + L + W + D$	$\lambda + \alpha + \beta + \gamma + (\gamma\alpha)$	12,6	7
$W \times D + L + W + D$	$\lambda + \alpha + \beta + \gamma + (\gamma\beta)$	23	9
$L \times W + L \times D + L + W + D$	$\lambda + \alpha + \beta + \gamma + (\beta\alpha) + (\gamma\alpha)$	5,8	4
$L \times W + W \times D + L + W + D$	$\lambda + \alpha + \beta + \gamma + (\beta\alpha) + (\gamma\beta)$	13,8	6
$L \times D + W \times D + L + W + D$	$\lambda + \alpha + \beta + \gamma + (\gamma\alpha) + (\gamma\beta)$	10,8	6
$L \times W + L \times D + W \times D + L + W + D$	$\lambda + \alpha + \beta + \gamma + (\beta\alpha) + (\gamma\alpha) + (\gamma\beta)$	2,4	3

Dobrze dopasowanym do danych i jednocześnie najprostszym okazał się model dwuczynnikowy obejmujący interakcję *Wiek* i *Pragnienia posiadania dzieci w przyszłości* ($L \times D + L + W + D$). Interpretacja graficzna tej sytuacji pokazana jest na rys. 13.



Rys. 13. Wykres logitów przy interakcji *Wiek* i *Pragnienie posiadania dzieci*.

Wykres pokazuje cztery krzywe logitowe stosowania antykoncepcji w zależności od wieku dla grup określonych przez wykształcenie i pragnienie posiadania dzieci. Krzywe są oznaczone symbolami W i N dla wyższego i niższego wykształcenia oraz T i N dla pragnienia posiadania dzieci w przyszłości. Najniższa krzywa oznaczona przez NT odpowiada kobietom o niższym wykształceniu, które chcą mieć dzieci, i pokazuje niewielki wzrost w stosowaniu antykoncepcji. Następna krzywa oznaczona przez WT dotyczy kobiet z wyższym wykształceniem, które również chcą mieć dzieci. Ta krzywa jest równoległa do

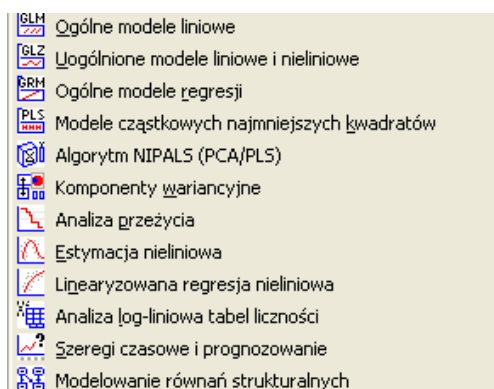


poprzedniej, ponieważ efekt edukacji jest addytywny względem wieku. Stała różnica między tymi dwiema krzywymi oznacza wzrost w ilorazie szans przy przejściu od niższego do wyższego wykształcenia. Trzecia krzywa, oznaczona przez NN dotyczy kobiet z niższym wykształceniem, które nie chcą mieć dzieci w przyszłości. To, że kobieta nie pragnie w przyszłości mieć dzieci, manifestuje się przesunięciem krzywych. Ten efekt wzrasta ostro z wiekiem, osiągając iloraz szans cztery dla 40-49 lat. Czwarta krzywa, oznaczona symbolem WN, odpowiada kobietom z wyższym wykształceniem, które nie chcą w przyszłości mieć dzieci. Odległość między tą krzywą a poprzednią to efekt wykształcenia, który jest taki sam niezależnie od tego, czy kobiety chcą czy nie mieć w przyszłości dzieci i jest też niezależny od wieku.

Czy to oznacza, że możemy skończyć nasze poszukiwania odpowiedniego modelu? Niekoniecznie. Nasz model nie uwzględnił (jak podpowiadała analiza korespondencji) interakcji efektu *Wykształcenia*. W stwierdzeniu, czy uwzględnienie tych interakcji jest konieczne, pomoże nam analiza log-liniowa.

Wykorzystanie analizy log-liniowej

Modelując interesujące nas powiązania, moglibyśmy postąpić jeszcze inaczej. Statystycy opracowali specjalną analizę wielowymiarowych tabel wielodzzielczych, która umożliwia testowanie istotności wpływu różnych czynników (ujętych w tabeli) i ich interakcji. Poszukiwanie odpowiedniego modelu w wielodzzielczej (większej niż dwudzzielcza) tabeli liczebności jest często zadaniem trudnym. Moduł *Analiza log-liniowa* daje wiele możliwości ułatwiających takie poszukiwania. Zatem zebrane dane przeanalizujemy również za pomocą analizy log-liniowej.



Rys. 14. Menu wyboru analizy log-liniowej.

Najważniejsze wyniki zawierają poniższe dwa arkusze wyników.



k czynn.	Wyniki dop. wszystkich interakcji k czynników (Liczność Jednoczesne testy, że wszystkie interakcje k czynników są jednocześnie równe 0.				
	Stopnie swobody	chi-kwad naj.wiar	Prawdop. p	chi-kwad Pearsona	Prawdop. p
1	6	607,8530	0,000000	685,5132	0,000000
2	12	539,8292	0,000000	631,4675	0,000000
3	10	27,9174	0,001861	26,8311	0,002769
4	3	2,4407	0,486095	2,5039	0,474579

Rys. 15. Arkusz wyników wszystkich interakcji.

Z pierwszego arkusza wynika, że analizowane powinny być modele uwzględniające oprócz antykoncepcji interakcje, co najwyżej 2-czynnikowe pozostałych czynników.

Drugi arkusz to testy wszystkich modeli brzegowych i cząstkowych. Możemy zobaczyć, które z dwuwymiarowych i trójwymiarowe zależności są istotne.

Efekt	Testy związku brzegowego i cząstkowego				
	Stopnie swobody	Zw.cząst chi-kwad	Zw.cząst p	Z.brzeg. chi-kwad	Z.brzeg. p
1- Dzieci	1	70,4863	0,000000	70,4863	0,000000
2 - Wiek	3	225,3218	0,000000	225,3218	0,000000
3 - Wykształcenie	1	90,2798	0,000000	90,2798	0,000000
4 - Antykoncepcja	1	221,7641	0,000000	221,7641	0,000000
12	3	126,6669	0,000000	195,4729	0,000000
13	1	1,1883	0,275680	29,3044	0,000000
14	1	50,6023	0,000000	91,0768	0,000000
23	3	186,8456	0,000000	209,4528	0,000000
24	3	43,2347	0,000000	78,2013	0,000000
34	4	6,4827	0,010893	0,7587	0,383746
123	3	0,9383	0,816172	1,4424	0,695616
124	3	10,7296	0,013282	15,8666	0,001208
134	1	3,4263	0,064168	5,9094	0,015060
234	3	8,3333	0,039603	7,1320	0,067907

Rys. 16. Arkusz wyników testów związku brzegowego i cząstkowego.

Zależność cząstkowa informuje o tym, czy odpowiednia interakcja jest istotna, gdy wszystkie inne efekty tego samego stopnia są już w modelu. Spójrzmy na przykład na Efekt 14. Efekt ten reprezentuje powiązanie lub interakcję między czynnikami 1-Dzieci i 4-Antykoncepcja. Gdy usuniemy ten efekt z modelu z wszystkimi innymi zależnościami dwuwymiarowymi, różnica w wartości chi-kwadrat (największej wiarygodności) wynosi 50,6 z 1 stopniem swobody. Wartość ta jest istotna na poziomie $p = 0,0000$. Dlatego dopasowanie modelu staje się istotnie gorsze, gdy eliminujemy tę dwuwymiarową interakcję z modelu; zatem pozostawiamy ją.

Test zależności brzegowej efektu 14 odnosi się do różnicy między modelem nieuwzględniającym żadnych interakcji dwuwymiarowych a modelem, który zawiera interakcję 14 (i żadnych innych interakcji dwuwymiarowych). Jak widać, dopasowanie modelu poprawia się istotnie, gdy dodajemy zależność między czynnikami 1-Dzieci i 4-Antykoncepcja (chi-kwadrat = 91,08, $df = 1$, $p = 0,0000$).



Wybór efektów do modelu

Przede wszystkim interesują nas jednak czynniki, które są związane z 4-Antykoncepcją. Z arkusza zawierającego wyniki testu wszystkich modeli brzegowych i cząstkowych wnioskujemy, że do modelu powinniśmy włączyć efekty: 14, 24, 124 (czcionka pogrubiona na rys. 16).

Efekt 134 (wyróżniony kursywą w arkuszu na rys. 16) to zależność między 1-Dzieci i 3-Wykształcenie i 4-Antykoncepcją, która nie jest istotna, gdy oceniamy ją z wszystkimi innymi zależnościami, może być zatem wytłumaczona przez inne efekty. Zatem nie włączamy jej teraz do modelu.

Najwięcej kłopotu sprawiają efekty związane z wiekiem 34, 234 (czcionka przekreślona na rys. 16). Nie są one istotne same, ale w obecności innych efektów tego samego rzędu stają się istotne. Taka zmienna nosi nazwę zmiennej zakłócającej i powinna być włączona do modelu. Zatem efekt 234 to zależność 2-Wieku, 3-Wykształcenia i 4-Antykoncepcją, którą włączamy do modelu. Potwierdziła ona, że najlepszym modelem jest model dwuczynnikowy uwzględniający wpływ interakcji wieku i wykształcenia.

Dalszą analizę modelu kontynuujemy z zastosowaniem ogólnego modelu liniowego. Analizujemy model $\lambda + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik}$

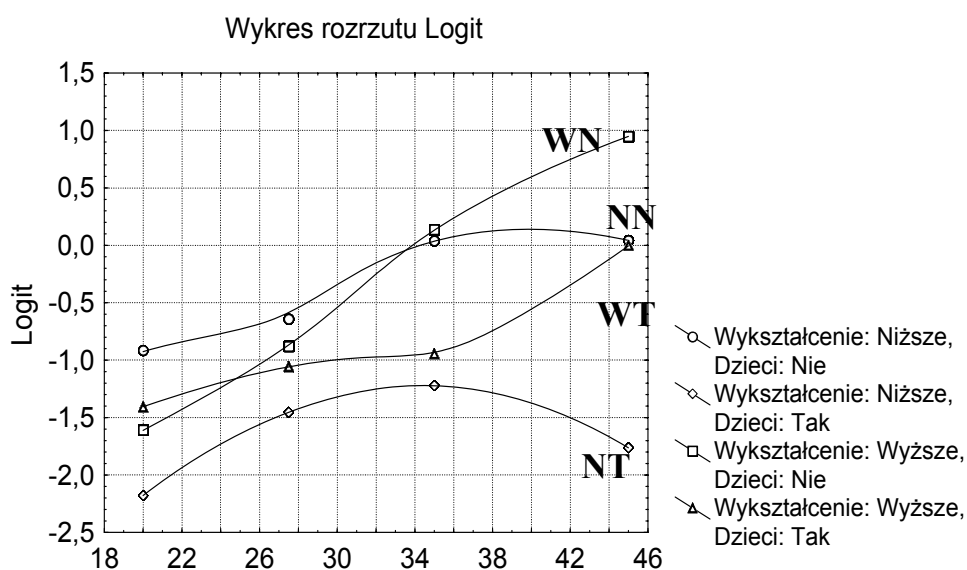
gdzie: λ – logit prawdopodobieństwa stosowania antykoncepcji dla kobiet poniżej 25 lat i które chcą mieć dzieci w przyszłości (poziom odniesienia),

α_i – efekt netto grup wiekowych w porównaniu do grupy <25 w tej samej kategorii pragnienia posiadania dzieci,

β_j – efekt netto „niechcących” mieć dzieci w stosunku do tych, które „chcą”,

γ_k – efekt netto wykształcenia.

Interpretacja graficzna tej sytuacji pokazana jest na rys. 17.



Rys. 17. Wykres logitów dla analizy wieloczynnikowej (*Wiek, Wykształcenie i Pragnienie posiadania dzieci*).



Rysunek ten pokazuje dopasowane wartości bazujące na bardziej złożonym modelu. Stosowanie antykoncepcji dla pragnących mieć dzieci wzrasta nieznacznie do lat 35 i wtedy opada dla mniej wykształconych, ale nadal wzrasta dla bardziej wykształconych. Stosowanie antykoncepcji dla niepragnących mieć dzieci w przyszłości wzrasta ostro z wiekiem do lat 35, i wtedy wyrównuje się dla mniej wykształconych, ale nadal wzrasta dla bardziej wykształconych. Rysunek pokazuje, że efekt niechcących w przyszłości mieć dzieci wzrasta z wiekiem i wydaje się, że dla obu edukacyjnych grup w ten sam sposób (popatrz na odległość między krzywymi NN a NT i między krzywymi WN a WT). Efekt edukacji jest lepiej widoczny przy 40-49 latach niż we wcześniejszych latach i też wydaje się nieznacznie większy dla kobiet, które pragną mieć dzieci niż dla tych, którzy nie pragną (por. odległości między krzywymi NT a WT i między krzywymi NN a WN). Zatem dołączenie interakcji *Wykształcenia* w pełni tłumaczy istniejące powiązania.

Podsumowanie i wnioski

Przedstawione w artykule metody analizy oraz ich wyniki miały przede wszystkim na celu zilustrowanie różnych technik opracowania danych nominalnych. Ponadto chodziło o zaprezentowanie wybranych narzędzi wspomaganie analizy danych w programie *STATISTICA 8*. Przeprowadzona analiza dotyczyła co prawda zagadnień z zakresu medycyny, ale wydaje się, że może również zostać z powodzeniem wykorzystana w przypadku analizy zagadnień pochodzących z innych dziedzin badań empirycznych.

W charakterze podsumowania można przedstawić poniższe wnioski końcowe:

Wnioski merytoryczne związane z analizowanymi danymi

- ◆ Widoczny jest istotny wpływ wieku objawiający się widocznym wzrostem stosowania antykoncepcji wraz ze wzrostem wieku.
- ◆ Wzrost ten zależy od pragnienia urodzenia dzieci i wykształcenia.
- ◆ Dla grupy niepragnących w przyszłości mieć dzieci wzrost ten jest ostry, dochodząc do ilorazu szans równego cztery w grupie 40–49.
- ◆ Dla grupy pragnącej mieć dzieci wzrost jest o wiele łagodniejszy do wieku 30–39. Dalszy przebieg zależy od wykształcenia. Dla osób z wykształceniem wyższym następuje dalszy wzrost. Natomiast dla osób z wykształceniem niższym następuje nieznaczny spadek.
- ◆ Dla osób z wyższym wykształceniem widoczne jest zwiększenie stosowania antykoncepcji, zwłaszcza w najstarszej grupie wiekowej.

Wnioski ogólne

- ◆ W przypadku zagadnień, w których występują różnorakie powiązania pomiędzy badanymi zmiennymi, warto stosować różne techniki modelowania. Okazało się, że nie wystarcza tradycyjne chi-kwadrat. Interesujących powiązań nie otrzymamy za pomocą prostych analiz tabel wielodzzielczych.



- ◆ Przedstawione w artykule sposoby opracowania tych konkretnych danych mogą zostać z powodzeniem wykorzystane również w przypadku danych pochodzących z innych badań.
- ◆ Nowa wersja programu *STATISTICA* pozwala znakomicie wspomagać stosowanie różnych technik statystycznej analizy danych.

Literatura

1. Amemiya T., *Some theorems in the linear probability model*, „International Economic Review”, 18, 645–650, 1977.
2. Agresti A., *Categorical Data Analysis*, Wiley, New York, 1990.
3. Christensen R., *Log-linear Models and Logistic Regression*, Springer-Verlag, New York, 1990.
4. Maddala G. S., *Ekonometria*, Wydawnictwo Naukowe PWN, 2006.
5. Stanisław A., *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny*, StatSoft Polska, Kraków, 2007.