Uniwersytet Warszawski
Wydział Filozofii i Socjologii
Instytut Socjologii

Dariusz Przybysz

# Modele logarytmiczno–liniowe dla zmiennych porządkowych

## Analiza tablic ruchliwości i danych panelowych

Praca doktorska
napisana pod kierunkiem
prof. dra hab. Grzegorza Lissowskiego

Warszawa, wrzesień 2009 r.

# Spis treści

**Wprowadzenie** **1**

**1 Modele logarytmiczno–liniowe dla zmiennych nominalnych** **6**

1.1 Formułowanie hipotez dotyczących rozkładu łącznego zmiennych . . . 7

1.1.1 Hipotezy dla dwóch zmiennych nominalnych . . . . . 8

1.1.2 Hipotezy dla trzech zmiennych nominalnych . . . . . 13

1.1.3 Hipotezy dla większej liczby zmiennych . . . . . 36

1.2 Parametryzacja modelu logarytmiczno–liniowego . . . . . 36

1.2.1 Wersja multiplikatywna i addytywna modelu . . . . . 37

1.2.2 Różne wersje parametryzacji . . . . . 38

1.2.3 Modele hierarchiczne i niehierarchiczne . . . . . 50

1.3 Estymacja i weryfikacja modeli . . . . . 51

1.3.1 Estymacja . . . . . 52

1.3.2 Iloraz wiarygodności, statystki $G^2$ oraz $X^2$ . . . . . 54

1.3.3 Inne metody oceny dopasowania modelu . . . . . 62

1.4 Przykład analizy empirycznej . . . . . 70

**2 Modele logarytmiczno–liniowe dla zmiennych porządkowych** **75**

2.1 Lokalne stosunki szans i ich własności . . . . . 76

2.2 Niemniejszość stochastyczna . . . . . 77

2.3 Modele dla dwóch zmiennych . . . . . 79

2.3.1 Model jednakowej interakcji (UA) . . . . . 80

2.3.2 Model efektu wierszowego (R) . . . . . 87

2.3.3 Model efektu wierszowo–kolumnowego (RC1) . . . . . 97

2.3.4 Model logarytmiczno–multiplikatywny (RC2) . . . . . 102

2.4 Modele dla trzech zmiennych . . . . . 109

2.4.1 Modelowanie interakcji drugiego rzędu . . . . . 110

2.4.2 Modelowanie interakcji trzeciego rzędu . . . . . 124

**3 Modele dla tablic ruchliwości i danych panelowych 160**
3.1 Przykłady tabel o takich samych kategoriach zmiennych . . . . . 161
3.2 Modele dla dwóch zmiennych o takich samych kategoriach . . . . . 176
3.2.1 Niezależność i quasi–niezależność stochastyczna . . . . . 176
3.2.2 Model symetrii i quasi–symetrii . . . . . 190
3.2.3 Modele jednakowej interakcji i wierszowo–kolumnowe . . . . . 197
3.2.4 Modele dystansu i przekraczania barier . . . . . 212
3.2.5 Podsumowanie omówionych modeli . . . . . 234
3.3 Modelowanie asymetrii . . . . . 236
3.4 Modele dla większej liczby zmiennych . . . . . 248
3.4.1 Modelowanie warunkowej zależności dla tablic ruchliwości . . . 249
3.4.2 Analiza rozkładu trzech zmiennych o takich samych kategoriach 256
3.4.3 Zmiany w rozkładzie łącznym zmiennych — dane panelowe . . 259

**Podsumowanie 264**

**A Opis zbiorów danych, dodatkowe ilustracje i formuły 267**
A.1 Informacje o zbiorach danych . . . . . 267
A.2 Twierdzenie o agregacji (collapsibility theorem) . . . . . 269
A.3 Ilustracja do modelu quasi–niezależności . . . . . 270
A.4 Dodatkowe formuły dla wybranych modeli . . . . . 271
A.4.1 Model quasi–niezależności QN . . . . . 271
A.4.2 Model quasi–niezależności QhN . . . . . 273
A.4.3 Model QUA . . . . . 273
A.4.4 Model Q(R=C)1 . . . . . 275
A.4.5 Modele QD i QDS . . . . . 277
A.4.6 Model QFD . . . . . 278
A.5 Równoważność modeli dla tablicy o wymiarach 4 x 4 . . . . . 278
A.5.1 Modele QS i QDS . . . . . 278
A.5.2 Modele QCP i QFD . . . . . 278

**Bibliografia 280**

# Wprowadzenie

Analiza statystyczna zmiennych o charakterze porządkowym wydaje się szczególnie istotna z punktu widzenia socjologa. W badaniach kwestionariuszowych większość pytań — szczególnie dotyczących opinii respondenta w rozmaitych kwestiach — formułuje się w ten sposób, że uzyskiwane informacje mierzone są na skalach „słabych", tj. skali nominalnej lub porządkowej. Na ogół wynika to ze specyfiki badanych cech jak też braku wystarczającego uzasadnienia teoretycznego, które pozwalałoby na pomiar tych cech na skali mocniejszej[1] np. interwałowej bądź ilorazowej. O ile pewne cechy dają się mierzyć w ten sposób (przykładowo wiek, wzrost) to trudno byłoby zbudować narzędzie pomiarowe do precyzyjnego określenia różnych postaw, np. tolerancji wobec obcokrajowców, akceptacji postaw egalitarnych itd.

W związku z tym, powstaje problem adekwatnej analizy statystycznej danych tego typu. Z jednej strony metoda powinna być dopuszczalna ze względu na skale, na jakiej mierzone są zmienne. Przykładowo, jeśli posiadamy informację o wykształceniu respondentów i wyodrębniono kategorie *podstawowe*, *średnie*, *wyższe*, liczenie średniej w celu opisu tej zbiorowości jest niedopuszczalne tj. nie ma teoretycznego uzasadnienia. Nie wiadomo, jakie wartości liczbowe powinno przypisać się kolejnym kategoriom, nie wiemy, czy odległość pomiędzy kategoriami *podstawowe–średnie* jest taka sama, mniejsza, czy też większa aniżeli w przypadku kategorii *średnie–wyższe*. W konsekwencji, w zależności od tego, jakie wartości zostaną przypisane może się okazać, że w badanej zbiorowości kobiety są „średnio" lepiej wykształcone od mężczyzn, bądź można sformułować wniosek przeciwny: średnia wykształcenia jest wyższa wśród mężczyzn. Z drugiej strony, pożądane jest wykorzystanie wszystkich informacji, jakie ta zmienna zawiera. Przykładowo: inną charakterystyką może być wskazanie kategorii najczęściej występującej (modalnej). Jej użycie jest całkowicie dopuszczalne, co więcej może to być informacja przydatna do celów opisu zbiorowości. Należy jednak zauważyć, że charakterystyka ta abstrahuje od tego, czy poszczególne kategorie są, czy też nie są uporządkowane. Powyższe kategorie wykształcenia tworzą pewną hie-

[1] Szersze omówienie kwestii pomiaru i skal pomiarowych można znaleźć np. w Lissowski i inni (2008).

rarchię, tj. nie można posiadać wykształcenia wyższego nie kończąc wcześniej szkoły średniej.

Powyższy przykład pokazuje, że analizując dane z trzeba zwracać uwagę na to, czy metoda jest dopuszczalna, ponadto warto wykorzystywać te metody, które obejmują wszystkie informacje, jakie zawiera zmienna ze względu na skalę, na której jest mierzona. Średnia jest dopuszczalna dla skal interwałowych, modalna dla skali nominalnej, czyli w praktyce każdej zmiennej. Nie znaczy to, że do opisu zmiennej porządkowej nie należy wykorzystywać metod dopuszczalnych dla zmiennych nominalnych. Niemniej warto wzbogacić opis statystyczny takiej zmiennej poszukując metod i parametrów, które wykorzystują informację o niearbitralnym uporządkowaniu kategorii, np. wskazać wartość minimalną, maksymalną bądź medianę tej zmiennej.

W ubiegłym stuleciu sformułowano wiele metod adekwatnych dla zmiennych porządkowych, (m.in. Kendall 1948, Kruskal 1958, Goodman i Kruskal 1959, Hildebrandt i inni 1977, Lissowski 1978). Jedną z nich są modele logarytmiczno–liniowe. Metoda ta zaczęła się rozwijać w latach 60–tych XX wieku (Birch 1963) i jako jej głównego autora należy wskazać Leo Goodmana (1963, 1970, 1971). Od tego czasu powstało wiele prac, które przyczyniły się do rozwoju tej metody, bądź popularyzowały ją wśród badaczy (m. in. Bishop i inni 1975, Fienberg 1980, Haberman, 1974a, 1978, 1979, Knoke i Burke 1980, Andersen 1980, Fingleton 1984, Liu i Agresti 2005, Goodman 2007). W najbardziej ogólnym zarysie metoda ta służy do analizy rozkładu łącznego dwóch bądź większej liczby zmiennych. W jej ramach formułuje się hipotezy dotyczące rozkładu poszczególnych zmiennych i związków pomiędzy nimi. Poszczególne modele zwykło się opisywać za pomocą parametrów o możliwie dogodnej interpretacji.

Trzeba zaznaczyć, że początkowo metoda ta dotyczyła zmiennych nominalnych. Pierwsze prace wskazujące na możliwości uwzględnienia porządkowego charakteru zmiennych pojawiły się w latach 70–tych (Goodman 1979*b*, Haberman 1974*a*). Modele tego typu były w kolejnych latach niejednokrotnie prezentowane i rozwijane (Agresti 1984, Clogg 1982*a*, Ishii-Kuntz 1994).

Pewną szczególną klasę modeli logarytmiczno–liniowych stanowią dane, w których kategorie analizowanych zmiennych ściśle ze sobą korespondują. Przykładem może być tablica ruchliwości edukacyjnej, w której kategorie wykształcenia syna zestawione są z kategoriami wykształcenia jego ojca. Podobną strukturę mają dane panelowe, tablice opisujące wzory zawierania małżeństw, gdzie zestawia się tę samą cechę dla męża i żony (np. zawód). Specyficzna struktura danych tego typu pozwala na formułowanie odrębnych modeli (m.in. Goodman 1972a, Haberman 1979, Hout 1983, Hagenaars 1990, Yamaguchi 1990, Lawal 2003, 2004). Warto już w tym miejscu

podkreślić, że zagadnienia dotyczące ruchliwości społecznej miały istotny wpływ na rozwój modelowania logarytmiczno-liniowego. Pytania badawcze, potrzeba pomiaru pewnych zjawisk społecznych związanych z ruchliwością, np. „dziedziczenia" pozycji przyczyniły się do formułowania nowych modeli[2].

Wykorzystanie modeli logarytmiczno–liniowych w polskiej socjologii zapoczątkował Michał Pohoski (1983). Prezentację samej metody skierowaną do polskiego Czytelnika zawiera artykuł Grzegorza Lissowskiego (1984). Można wskazać wiele prac napisanych w języku polskim bądź prezentujących samą metodę bądź wykorzystujących ją jako narzędzie do analizy danych, m. in. Misztal(1982, 1990), Pohoski (1983, 1991), Nawojczyk i McCutcheon (1996), Domański (1989, 1996, 2004), Przybysz (2003), Domański i Przybysz(2007), Domański i inni (2008), jak również prace napisane przez polskich socjologów w języku angielskim, m.in. Kutyłowski (1989, 1994), Mach (2002, 2004), Domański (2007*a*). Warto jednak nadmienić, że w pracach tych stosunkowo rzadko uwzględniano porządkowy charakter zmiennych.

Zanim przejdę do omówienia planu tej pracy, warto podkreślić co odróżnia ją od innych publikacji dotyczących tej metody analizy danych. Będzie ona dotyczyć będzie przede wszystkim możliwości uwzględnienia zmiennych porządkowych w modelowaniu logarytmiczno–liniowym. Szczególnie dużo miejsca poświęcone zostanie zasygnalizowanej powyżej kwestii analizy danych panelowych i tablic ruchliwości, czyli sytuacji, gdy kategorie zmiennych są identyczne.

Model logarytmiczno–liniowy dość często jest postrzegany jako szczególny typ modelu liniowego (Dobson 2002, Agresti 2002), co pozwala dostrzec pewne analogie pomiędzy tą metodą a innymi powszechnie stosowanymi, np. analizą wariancji czy regresją liniową. W takim ujęciu model logarytmiczno–liniowy można przedstawić następująco: zmienna zależną jest logarytm liczebności rozkładu łącznego i jest on funkcją parametrów związanych z poszczególnymi kategoriami zmiennych i ich kombinacjami. Parametry takiego równania (lub jej multiplikatywnego odpowiednika, gdzie parametry nie są dodawane a mnożone przez siebie) pozwalają opisywać rozkład, jak również formułować hipotezy dotyczące rozkładu. Praca ta — choć ujęcie takie również jest w niej przedstawione — przyjmuje inny sposób prezentacji modelowania logarytmiczno–liniowego. Omówienie każdego modelu rozpoczyna sformułowanie hipotezy za pomocą pojęć związanych w sposób szczególny z tą metodą: równomierności rozkładu, niezależności stochastycznej, bądź innych relacji pomiędzy poszczególnymi

[2]Warto poczynić w tym miejscu pewną dygresję historyczną. Uwzględnienie porządkowego charakteru zmiennych w modelowaniu logarytmiczno–liniowym po raz pierwszy miało miejsce właśnie w odniesieniu do tablic o takich samych kategoriach obydwu zmiennych (Goodman 1972*c*). Modele te nie mogły być jednak stosowane w odniesieniu do sytuacji, gdy kategorie analizowanych zmiennych różniły się od siebie. Te zostały sformułowane później w pracach przytoczonych powyżej.

prawdopodobieństwami, między innymi. tzw. stosunków szans. U podstaw takiej prezentacji leży przekonanie, że pozwala ona pełniej dostrzec własności poszczególnych modeli. Ujęcie takie jest rzadziej wykorzystywane w literaturze. Oczywiście było ono stosowane w wielu pracach, trzeba jednak przyznać, że np. w odniesieniu do tablic ruchliwości wykorzystywane było ono sporadycznie. Trudno wskazać na pracę, która modele logarytmiczno–liniowe różnego typu przedstawiałaby systematycznie z tej perspektywy[3]. Próbę taką podejmuję w niniejszej rozprawie.

Choć praca dotyczy głównie modelowania dla zmiennych porządkowych, to pierwszy rozdział omawia zastosowanie modeli logarytmiczno–liniowych dla zmiennych nominalnych. Z jednej strony, na tym przykładzie łatwiej wyjaśnić logikę formułowania modeli, z drugiej strony, modele te często traktuje się jako punkt odniesienia w stosunku do tych, które uwzględniają uporządkowanie kategorii zmiennych. Nacisk położony zostanie na prezentację hipotez związanych z przedstawianymi modelami. W pierwszej kolejności zaprezentowane będą hipotezy dla dwóch zmiennych, w dalszej części — dla trzech i większej liczby zmiennych. Rozdział ten obejmuje również ogólną prezentację kwestii parametryzacji, estymacji i weryfikacji modelu. Ta sama hipoteza może być przedstawiona w wersji parametrycznej na wiele różnych sposobów. Porównane zostaną ze sobą dwie najbardziej rozpowszechnione w literaturze metody parametryzacji: odchyleń multiplikatywnych (*effect coding*) oraz parametryzacja względem kategorii odniesienia (*dummy coding*). Dalsza część tego rozdziału dotyczyć będzie estymacji modelu metodą największej wiarygodności (*maximum likelihood estimation*), z wykorzystaniem przykładu szacowania rozkładu oczekiwanego dla wybranej, relatywnie prostej hipotezy. Jeśli chodzi o zagadnienie weryfikacji modelu dość szczegółowo omówiona zostanie kwestia wykorzystania statystyki $G^2$ opartej na tzw. ilorazie wiarygodności, jak również sposobu wyznaczenia liczby stopni swobody dla danego modelu. Na końcu tego rozdziału znaleźć można krótkie omówienie innych mierników dopasowania modelu do danych (m. in. indeksu rozbieżności, indeksu BIC).

Rozdział drugi dotyczy modeli, w których jedna bądź więcej zmiennych mierzona jest na skali porządkowej. Przedstawiony zostanie model jednakowej interakcji (*uniform association*), model wierszowy (*row effect model*) i model wierszowo–kolumnowy (*row–column effect model*) w dwóch wersjach. Dla lepszego zrozumienia tych modeli wykorzystane będą pojęcia niemniejszości stochastycznej i zależności regresyjnej. Dalsza część rozdziału będzie dotyczyć modeli dla większej liczby zmiennych. Na koń-

[3]Przykładowo, w odniesieniu do trzech zmiennych nominalnych należałoby wskazać pracę Michała Bojanowskiego (2003), dla tablic o takich samych kategoriach w pewnym zakresie również artykuł Lawala (2003)

cu przedstawiona zostanie propozycja Goodmana–Houta, która stanowi uogólnienie szerokiej klasy modeli.

Rozdział trzeci – najobszerniejszy – poświęcony jest analizie danych, w których kilka zmiennych ma takie same kategorie, w szczególności tablicom ruchliwości i danym panelowym. Prezentację rozpoczną przykłady danych o strukturze tego typu, odsetki za pomocą których można charakteryzować tego rodzaju tabele. W dalszej części zaprezentowane zostaną modele logarytmiczno-liniowe, jakie można stosować w odniesieniu do tabel dla dwóch zmiennych o takich samych kategoriach. Poza modelami, które pojawiły się w dwóch pierwszych rozdziałach, pojawią się modele specyficzne dla danych tego typu, między innymi model quasi–niezależności, model symetrii i modele wymagające porządkowego pomiaru zmiennych: modele dystansu i przekraczania barier. Następnie zaprezentowana zostanie kwestia uwzględnienia asymetrii w modelach tego typu. Prezentację kończą modele dla większej liczby zmiennych, z których co najmniej dwie mają takie same kategorie.

Warto podkreślić, że zarówno w drugim, jak i trzecim rozdziale, prezentacja modeli — podobnie jak w rozdziale pierwszym — będzie rozpoczynać się od sformułowania hipotezy. W drugiej kolejności przedstawiony zostanie przykład jej parametryzacji i interpretacji poszczególnych parametrów. Omawiana będzie również kwestia estymacji i liczby stopni poszczególnych modeli.

W pracy tej prezentacji modeli towarzyszyć będą liczne ilustracje oparte bądź na fikcyjnych rozkładach bądź na danych empirycznych z badań sondażowych. Warto podkreślić, że celem tych ilustracji było przede wszystkim ułatwienie zrozumienia poszczególnych modeli. Analizy te nie służą do wyciągania ogólnych wniosków w odniesieniu do opisywanych przez nie zjawisk, przykładowo praca ta nie ma ambicji do opisu kwestii ruchliwości edukacyjnej w Polsce. Podobnie nie są tu podejmowane próby dyskusji z hipotezami badawczymi formułowanymi na gruncie teorii ruchliwości społecznej bądź z najważniejszymi ustaleniami badań w tej dziedzinie. Cel tej pracy jest przede wszystkim metodologiczny i dobór przykładów został temu podporządkowany.

# Rozdział 1

# Modele logarytmiczno–liniowe dla zmiennych nominalnych

Modele logarytmiczno–liniowe służą do analizy rozkładu łącznego dwóch lub większej liczby zmiennych. Metoda ta pozwala na formułowanie hipotez dotyczących rozkładu zmiennych i związków pomiędzy nimi, np. niezależności stochastycznej. Ponadto, modele logarytmiczno–liniowe dają możliwość przedstawienia rozkładu łącznego zmiennych za pomocą niewielkiej liczby parametrów. Im prostsza jest hipoteza, tym mniej parametrów potrzebnych jest do opisu rozkładu. Parametry te opisują rozkłady poszczególnych zmiennych i związki pomiędzy nimi.

Jak zostało powiedziane we wstępie modele logarytmiczno–liniowe są szczególnie przydatne do analizy zmiennych jakościowych. Niektóre modele można zastosować do analizy zmiennych nominalnych, w innych zakłada się, że zmienne mierzone są na skali porządkowej. W tej części omówione zostaną modele logarytmiczno–liniowe dla zmiennych nominalnych.

W pierwszej kolejności przedstawione zostaną hipotezy dotyczące rozkładu łącznego dwóch zmiennych wykorzystywane w modelach logarytmiczno–liniowych. Następnie sformułowane zostaną hipotezy dotyczące trzech zmiennych i zasygnalizowana zostanie możliwość formułowania hipotez dla rozkładów o większej liczbie wymiarów. W dalszej kolejności zaprezentowane zostanie w jaki sposób rozkłady związane z tymi hipotezami są parametryzowane w modelach logarytmiczno–liniowych. Przedstawione zostaną różne rodzaje parametryzacji i omówione zostaną różnice w interpretacji parametrów przy zastosowaniu każdej z nich. Następnie przedstawiona zostanie kwestia estymacji w modelu w oparciu o dane z próby, a także weryfikacji statystycznej hipotez związanych z poszczególnymi modelami. Modele przedstawione w tym rozdziale wykorzystać można również do analizy zmiennych mierzonych na silniejszej

skali pomiaru np. na skali porządkowej. Jednak w takim przypadku informacja o uporządkowaniu kategorii zmiennej nie jest w tych modelach wykorzystana.

## 1.1 Formułowanie hipotez dotyczących rozkładu łącznego zmiennych

Wiele statystycznych metod analizy danych opiera się na formułowaniu hipotez dotyczących badanego zjawiska i konfirmacji tych hipotez za pomocą testów empirycznych. Hipotezy te mogą wynikać z teorii, przypuszczeń i intuicji badacza. U podłoża modeli logarytmiczno–liniowych leżą hipotezy dotyczące rozkładu łącznego zmiennych i — co się z tym wiąże — rodzaju związku pomiędzy zmiennymi. W Tabeli 1.1 przedstawiony został rozkład łączny dwóch zmiennych X i Y.

Tabela 1.1: Rozkład łączny dwóch zmiennych $X$ i $Y$

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $\dots$ | $y_c$ | $\Sigma$ |
|---|---|---|---|---|---|---|
| $x_1$ | $\pi_{11}^{XY}$ | $\pi_{12}^{XY}$ | $\pi_{13}^{XY}$ | | $\pi_{1c}^{XY}$ | $\pi_1^X$ |
| $x_2$ | $\pi_{21}^{XY}$ | $\pi_{22}^{XY}$ | $\pi_{23}^{XY}$ | | $\pi_{2c}^{XY}$ | $\pi_2^X$ |
| $x_3$ | $\pi_{31}^{XY}$ | $\pi_{32}^{XY}$ | $\pi_{33}^{XY}$ | | $\pi_{3c}^{XY}$ | $\pi_3^X$ |
| $\vdots$ | | | | | | |
| $x_r$ | $\pi_{r1}^{XY}$ | $\pi_{r2}^{XY}$ | $\pi_{r3}^{XY}$ | | $\pi_{rc}^{XY}$ | $\pi_r^X$ |
| $\Sigma$ | $\pi_1^Y$ | $\pi_2^Y$ | $\pi_3^Y$ | | $\pi_c^Y$ | 1 |

Wprowadzone w tej tabeli oznaczenia będą stosowane w dalszej części pracy. Prawdopodobieństwo, że zmienna $X$ przyjmie wartość $x_i$ oznaczane będą $\pi_i^X = P(X = x_i)$ gdzie $i$=1, 2,..., $r$. Analogicznie przez $\pi_j^Y = P(Y = y_j)$ oznaczane będzie prawdopodobieństwo, że zmienna $Y$ przyjmie wartość $y_j$ gdzie $j$=1, 2,..., $c$. W przykładach tabelarycznych kategorie zmiennej $X$ będą występowały w wierszach, a kategorie zmiennej $Y$ w kolumnach. Przez $\pi_{ij}^{XY}$ oznaczać będziemy prawdopodobieństwa rozkładu łącznego dla kategorii $i-$tej zmiennej $X$ oraz kategorii $j-$tej zmiennej $Y$. Rozkłady brzegowe powstają przez zsumowanie prawdopodobieństw rozkładu łącznego po odpowiednim indeksie tj.

$$\pi_i^X = \sum_{j=1}^{c} \pi_{ij}^{XY} \qquad \text{oraz} \qquad \pi_j^Y = \sum_{i=1}^{r} \pi_{ij}^{XY}. \tag{1.1}$$

Poszczególne modele określają jaki jest rozkład prawdopodobieństwa dla danej hipotezy. Oczywiście, możliwe jest sformułowanie nieskończenie wielu hipotez dotyczą-

cych rozkładu łącznego zmiennych. W tabelach 1.2a i 1.2b przedstawione zostały przykładowe rozkłady łączne zmiennych X oraz Y. Hipotezy te — tak jak w podanym przykładzie — mogą być do siebie na tyle podobne, że ich rozróżnianie może być trudne do uzasadnienia. Choć każdy z tych rozkładów może adekwatnie opisywać związek pomiędzy pewnymi zmiennymi będącymi przedmiotem zainteresowania badacza, to bardziej interesujące wydaje się wyszczególnienie hipotez bardziej ogólnych, reprezentujących rozkłady pewnego typu.

Tabela 1.2: Przykładowe rozkłady łączne dwóch zmiennych

| Tabela 1.2a | | |
|---|---|---|
| X\Y | $y_1$ | $y_2$ |
| $x_1$ | 0,2 | 0,3 |
| $x_2$ | 0,1 | 0,4 |

| Tabela 1.2b | | |
|---|---|---|
| X\Y | $y_1$ | $y_2$ |
| $x_1$ | 0,19 | 0,30 |
| $x_2$ | 0,20 | 0,31 |

W modelach logarytmiczno–liniowych formułuje się hipotezy opisujące rozkłady o pewnej, szczególnej strukturze. Wiele z tych rozkładów daje się opisać za pomocą pojęć równomierności rozkładu i niezależności stochastycznej, warunkowej równomierności i warunkowej niezależności stochastycznej (tj. niezależności dwóch zmiennych w podzbiorowościach wyodrębnionych ze względu na wartości trzeciej zmiennej). Hipotezy dające się sformułować za pomocą tych pojęć nazywamy *hipotezami elementarnymi*. Niektóre hipotezy formułowane na gruncie modeli logarytmiczno-liniowych mają bardziej złożony charakter i do ich sformułowania wykorzystywane jest pojęcie *stosunku szans*, które zostanie wprowadzone w dalszej części tego rozdziału. W tym rozdziale przedstawione zostaną hipotezy formułowane dla zmiennych nominalnych, w pierwszej kolejności dla dwóch a następnie dla większej liczby zmiennych.

### 1.1.1 Hipotezy dla dwóch zmiennych nominalnych

Poniżej przedstawione zostaną hipotezy wykorzystywane do logarytmiczno–liniowego modelowania rozkładu łącznego dwóch zmiennych nominalnych. W przypadku dwóch zmiennych mamy do czynienia wyłącznie z hipotezami elementarnymi. Będą one prezentowane w kolejności od najprostszej do bardziej skomplikowanych. Poszczególne modele będą ilustrowane tabelarycznie, aby uwidocznić szczególne cechy tych rozkładów. Zaprezentowana wcześniej tabela 1.1 przedstawia sytuację dowolnego rozkładu dwóch zmiennych, nie zakłada się, że za rozkładem tym stoi jakakolwiek hipoteza. Prezentując tabelarycznie różne hipotezy będę starał się pokazać, ze rozkład zgodny z tymi hipotezami da się przedstawić za pomocą pewnej uproszczonej struktu-

ry. Dodatkowo poszczególne hipotezy zostaną sformułowane w odniesieniu do dwóch konkretnych zmiennych. Dwuwartościowa zmienna $X$ zdaje sprawę z tego, czy dana osoba uczestniczyła w wyborach (1. Nie, 2. Tak) , a zmienna $Y$ opisuje wykształcenie (1. podstawowe, 2. średnie, 3. wyższe).

### Hipoteza o równomierności rozkładu łącznego

Najprostszą sytuacją dla dwóch zmiennych jest równomierność rozkładu łącznego. Rozkład jest równomierny jeśli wszystkie prawdopodobieństwa rozkładu są sobie równe. Formalnie, zgodnie z tą hipotezą każde prawdopodobieństwo rozkładu łącznego można zapisać jako:

$$\pi_{ij}^{XY} = \frac{1}{r \cdot c} \qquad \text{dla każdej pary } i, j \tag{1.2}$$

W odniesieniu do zaproponowanego przykładu hipoteza ta oznacza, że prawdopodobieństwo wystąpienia każdej z sześciu kombinacji obydwu zmiennych (wykształcenia i uczestnictwa w wyborach) wynosi tyle samo. Przykładowo, prawdopodobieństwo, że osoba ma wykształcenie podstawowe i nie głosowała wynosi 1/6, tyle samo wynosi prawdopodobieństwo, że osoba ma wykształcenie średnie i brała udział w wyborach, itd. Jeśli przyjmie się, że prawdopodobieństwo danej wartości zmiennej definiuje odpowiedni odsetek[1] osób w zbiorowości oznacza to, że każda wymieniona wyżej podzbiorowość stanowi 1/6 całej zbiorowości.

Tabela 1.3: Rozkład łączny zgodny z hipotezą o równomierności

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $\dots$ | $y_c$ | $\Sigma$ |
|---|---|---|---|---|---|---|
| $x_1$ | $\frac{1}{r \cdot c}$ | $\frac{1}{r \cdot c}$ | $\frac{1}{r \cdot c}$ | | $\frac{1}{r \cdot c}$ | $\frac{1}{r}$ |
| $x_2$ | $\frac{1}{r \cdot c}$ | $\frac{1}{r \cdot c}$ | $\frac{1}{r \cdot c}$ | | $\frac{1}{r \cdot c}$ | $\frac{1}{r}$ |
| $x_3$ | $\frac{1}{r \cdot c}$ | $\frac{1}{r \cdot c}$ | $\frac{1}{r \cdot c}$ | | $\frac{1}{r \cdot c}$ | $\frac{1}{r}$ |
| $\vdots$ | | | | | | |
| $x_r$ | $\frac{1}{r \cdot c}$ | $\frac{1}{r \cdot c}$ | $\frac{1}{r \cdot c}$ | | $\frac{1}{r \cdot c}$ | $\frac{1}{r}$ |
| $\Sigma$ | $\frac{1}{c}$ | $\frac{1}{c}$ | $\frac{1}{c}$ | | $\frac{1}{c}$ | 1 |

Tabela 1.3 ilustruje rozkład łączny zgodny z powyższą hipotezą. Zauważmy, że w porównaniu do tabeli 1.1, prezentującej dowolny rozkład zmiennych, struktura

[1] Rozkład prawdopodobieństwa jest pojęciem bardziej ogólnym w stosunku do rozkładu częstości. Pamiętając o tym zastrzeżeniu interpretacja prawdopodobieństwa jako odpowiedniego odetka będzie wykorzystywana w dalszej części tej pracy, gdyż jest to pojęcie częściej używane w języku naturalnym.

rozkładu zgodna z hipotezą równomierności daje się przedstawić w bardzo prosty sposób. Do jednoznacznego wyznaczenia rozkładu łącznego zgodnego z tą hipotezą potrzebna jest jedynie informacja o liczbie kategorii obydwu zmiennych tj. $r$ i $c$.

### Hipoteza o równomierności warunkowej jednej ze zmiennych

Nieco bardziej skomplikowana hipoteza opisuje sytuację, w której rozkłady warunkowe jednej zmiennej wyróżnione ze względu na wartości drugiej zmiennej są równomierne. Jeżeli hipoteza głosi, że rozkład zmiennej $X$ jest równomierny względem zmiennej $Y$ wówczas wszystkie prawdopodobieństwa warunkowe zmiennej $X$ względem zmiennej $Y$ są takie same. Zgodnie z tą hipotezą:

$$\pi_{i(j)}^{X(Y)} = \frac{1}{r} \qquad \text{dla każdej pary } i,\, j. \tag{1.3}$$

Przez $\pi_{i(j)}^{X(Y)}$ oznaczone jest prawdopodobieństwo, że zmienna $X$ przyjmie $i-$tą wartość w podzbiorowości wyodrębnionej przez $j-$tą wartość zmiennej $Y$, tj. $\pi_{i(j)}^{X(Y)} = \pi_{ij}^{XY}/\pi_j^Y$. Wynika z tego, że prawdopodobieństwo rozkładu łącznego zgodne z tą hipotezą jest równe:

$$\pi_{ij}^{XY} = \frac{1}{r} \cdot \pi_j^Y \qquad \text{dla każdej pary } i,\, j. \tag{1.4}$$

Dla podanego przykładu powyższa hipoteza głosi, że dla każdej podzbiorowości wyróżnionej ze względu na wykształcenie, tj. wśród osób z wykształceniem podstawowym, średnim i wyższym połowa osób wzięła udział w głosowaniu.

Tabela 1.4: Rozkład zgodny z hipotezą o równomierności warunkowej $X$ względem $Y$

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $\dots$ | $y_c$ | $\Sigma$ |
|---|---|---|---|---|---|---|
| $x_1$ | $\frac{\pi_1^Y}{r}$ | $\frac{\pi_2^Y}{r}$ | $\frac{\pi_3^Y}{r}$ | | $\frac{\pi_c^Y}{r}$ | $\frac{1}{r}$ |
| $x_2$ | $\frac{\pi_1^Y}{r}$ | $\frac{\pi_2^Y}{r}$ | $\frac{\pi_3^Y}{r}$ | | $\frac{\pi_c^Y}{r}$ | $\frac{1}{r}$ |
| $x_3$ | $\frac{\pi_1^Y}{r}$ | $\frac{\pi_2^Y}{r}$ | $\frac{\pi_3^Y}{r}$ | | $\frac{\pi_c^Y}{r}$ | $\frac{1}{r}$ |
| $\vdots$ | | | | | | |
| $x_r$ | $\frac{\pi_1^Y}{r}$ | $\frac{\pi_2^Y}{r}$ | $\frac{\pi_3^Y}{r}$ | | $\frac{\pi_c^Y}{r}$ | $\frac{1}{r}$ |
| $\Sigma$ | $\pi_1^Y$ | $\pi_2^Y$ | $\pi_3^Y$ | | $\pi_c^Y$ | 1 |

Tabela 1.4 stanowi ilustrację rozkładu oczekiwanego zgodnego z modelem o równomierności warunkowej zmiennej X. Należy zauważyć, że istnieje wiele rozkładów zgodnych z powyższą hipotezą. Aby móc ten rozkład wyznaczyć jednoznacznie konieczna jest znajomość rozkładu zmiennej Y, przykładowo rozkład brzegowy. Zauważmy, że rozkład brzegowy tej zmiennej, jest taki sam jak rozkład $Y$ dla każdej kategorii

wyróżnionej ze względu na zmienną $X$, w tym sensie nie musi być to koniecznie informacja rozkładzie brzegowym[2].

Tabela 1.5: Rozkład zgodny z hipotezą o równomierności warunkowej $Y$ względem $X$

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $\dots$ | $y_c$ | $\Sigma$ |
|---|---|---|---|---|---|---|
| $x_1$ | $\frac{\pi_1^X}{c}$ | $\frac{\pi_1^X}{c}$ | $\frac{\pi_1^X}{c}$ | | $\frac{\pi_1^X}{c}$ | $\pi_1^X$ |
| $x_2$ | $\frac{\pi_2^X}{c}$ | $\frac{\pi_2^X}{c}$ | $\frac{\pi_2^X}{c}$ | | $\frac{\pi_2^X}{c}$ | $\pi_2^X$ |
| $x_3$ | $\frac{\pi_3^X}{c}$ | $\frac{\pi_3^X}{c}$ | $\frac{\pi_3^X}{c}$ | | $\frac{\pi_3^X}{c}$ | $\pi_3^X$ |
| $\vdots$ | | | | | | |
| $x_r$ | $\frac{\pi_r^X}{c}$ | $\frac{\pi_r^X}{c}$ | $\frac{\pi_r^X}{c}$ | | $\frac{\pi_r^X}{c}$ | $\pi_r^X$ |
| $\Sigma$ | $\frac{1}{c}$ | $\frac{1}{c}$ | $\frac{1}{c}$ | | $\frac{1}{c}$ | 1 |

Analogicznie można sformułować hipotezę o równomierności warunkowej zmiennej $Y$ względem zmiennej $X$. Wówczas zachodzi:

$$\pi_{(i)j}^{(X)Y} = \frac{1}{c} \qquad \text{dla każdej pary } i,\, j \tag{1.5}$$

oraz:

$$\pi_{ij}^{XY} = \frac{1}{c} \cdot \pi_i^X \qquad \text{dla każdej pary } i,\, j. \tag{1.6}$$

Tabela 1.5 ilustruje hipotezę o równomierności warunkowej zmiennej $Y$ względem $X$. Jak widać, do jednoznacznego wyznaczenia rozkładu zgodnego z tą hipotezą konieczne są informacje dotyczące rozkładu brzegowego zmiennej $X$. Dla omawianego przykładu hipoteza ta wskazuje, że zarówno wśród osób, które głosowały, jak też nie głosowały, 1/3 osób posiada wykształcenie podstawowe, 1/3 – wykształcenie średnie a pozostali wykształcenie wyższe. Hipoteza nie głosi natomiast jaki odsetek osób głosował, a jaki nie głosował.

### Hipoteza o niezależności stochastycznej zmiennych

Hipoteza ta opisuje sytuację, w której nie zakładamy nic na temat równomierności rozkładu żadnej ze zmiennych. Hipoteza ta głosi jedynie, że rozkłady warunkowe prawdopodobieństwa jednej zmiennej są takie same w podzbiorowościach wyodrębnionych przez wartości drugiej zmiennej i równe jej rozkładowi brzegowemu. Tj.

$$\pi_{i(j)}^{X(Y)} = \pi_i^X \qquad \text{dla każdej pary } i,\, j. \tag{1.7}$$

[2]Jak zobaczymy w dalszej części tego rozdziału, rozkład brzegowy wykorzystujemy w procesie estymacji rozkładu oczekiwanego na podstawie informacji z próby. Przy ogólnej prezentacji hipotez nie musi to być rozkład brzegowy, ale dla wygody będziemy wskazywali właśnie na ten rozkład, do niego będą też odwoływać się tabelaryczne ilustracje hipotez.

Warunek ten jest symetryczny, tak więc z 1.7 wynika również:

$$\pi_{(i)j}^{(X)Y} = \pi_j^Y \qquad \text{dla każdej pary } i,\, j. \tag{1.8}$$

Jak wiadomo, pomnożenie obydwu stron równania 1.7 przez $\pi_j^Y$ lub obydwu stron równania 1.8 przez $\pi_i^X$, pokazuje, że przy niezależności stochastycznej, prawdopodobieństwo rozkładu łącznego jest iloczynem odpowiednich prawdopodobieństw brzegowych:

$$\pi_{ij}^{XY} = \pi_i^X \cdot \pi_j^Y \qquad \text{dla każdej pary } i,\, j. \tag{1.9}$$

Tabela 1.6 jest ilustracją hipotezy o niezależności stochastycznej zmiennych. Jak widać do jednoznacznego wyznaczenia rozkładu związanego z tą hipotezą potrzebne są rozkłady brzegowe obydwu zmiennych. Odnosząc hipotezę o niezależności do wykształcenia i udziału w wyborach, wynika z niej, że w wśród osób głosujących i nie biorących udziału w wyborach rozkłady wykształcenia są identyczne. Podobnie, w każdej grupie wyróżnionej ze względu na wykształcenie, głosował taki sam odsetek osób. Hipoteza ta nie głosi jednak jakie są odsetki osób z wykształceniem podstawowym średnim lub wyższym, ani jaki jest odsetek osób głosujących.

Tabela 1.6: Rozkład zgodny z hipotezą o niezależności stochastycznej zmiennych

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $\dots$ | $y_c$ | $\Sigma$ |
|---|---|---|---|---|---|---|
| $x_1$ | $\pi_1^X \cdot \pi_1^Y$ | $\pi_1^X \cdot \pi_2^Y$ | $\pi_1^X \cdot \pi_3^Y$ | | $\pi_1^X \cdot \pi_c^Y$ | $\pi_1^X$ |
| $x_2$ | $\pi_2^X \cdot \pi_1^Y$ | $\pi_2^X \cdot \pi_2^Y$ | $\pi_2^X \cdot \pi_3^Y$ | | $\pi_2^X \cdot \pi_c^Y$ | $\pi_2^X$ |
| $x_3$ | $\pi_3^X \cdot \pi_1^Y$ | $\pi_3^X \cdot \pi_2^Y$ | $\pi_3^X \cdot \pi_3^Y$ | | $\pi_3^X \cdot \pi_c^Y$ | $\pi_3^X$ |
| $\vdots$ | | | | | | |
| $x_r$ | $\pi_r^X \cdot \pi_1^Y$ | $\pi_r^X \cdot \pi_2^Y$ | $\pi_r^X \cdot \pi_3^Y$ | | $\pi_r^X \cdot \pi_c^Y$ | $\pi_r^X$ |
| $\Sigma$ | $\pi_1^Y$ | $\pi_2^Y$ | $\pi_3^Y$ | | $\pi_c^Y$ | 1 |

### Podsumowanie hipotez dla dwóch zmiennych

Powyżej zostały przedstawione hipotezy jakie można sformułować dla dwóch zmiennych nominalnych. Tabela 1.7 stanowi podsumowanie powyższej prezentacji, zamieszczone zostały również oznaczenia hipotez, które będą używane w dalszej części tej pracy. Jak widać przedstawione hipotezy różnią się prostotą opisu. „Miarą" tej prostoty są ograniczenia nakładane na rozkład łączny obydwu zmiennych. Na przykład rozkład zgodny z hipotezą o równomierności jednej ze zmiennych jest prostszy od hipotezy o niezależności stochastycznej, gdyż zakłada dodatkowo, że rozkład jednej ze

zmiennych jest równomierny. Hipoteza o równomierności rozkładu zakłada, że rozkład obydwu zmiennych musi być równomierny. W tym sensie możliwe jest porównywanie ze sobą hipotez. Wyjątkiem są oczywiście hipotezy $[X]$ oraz $[Y]$. Są to hipotezy tego samego typu, dotyczą jednak innych zmiennych.

Zauważmy, że im mniej ograniczeń związanych jest z daną hipotezą, tym więcej informacji potrzebnych jest do jednoznacznego określenia rozkładu. Dla przykładu: hipoteza o niezależności stochastycznej zmiennych $X$ i $Y$ wymaga nie tylko informacji o rozkładzie $\{X\}$, które potrzebne są do określenia rozkładu zgodnego z prostszą hipotezą o równomierności warunkowej zmiennej $Y$, ale dodatkowo informacji o rozkładzie $\{Y\}$.

W tabeli — obok omówionych hipotez — wyszczególniona została dodatkowo sytuacja, gdy rozkład zmiennych jest dowolny. Do jednoznacznego wyznaczenia tego rozkładu potrzebnych jest najwięcej informacji. Z sytuacją taką nie jest związana żadna hipoteza, nie zakłada się nic na temat struktury rozkładu łącznego. Jak się okaże w dalszej części tego rozdziału wyszczególnienie tej sytuacji jest istotne z punktu widzenia modeli logarytmiczno–liniowych.

Tabela 1.7: Hipotezy dotyczące rozkładu łącznego dwóch zmiennych

| Hipoteza związana z rozkładem | Oznaczenie hipotezy | Rozkłady brzegowe konieczne do jednoznacznego określenia rozkładu zgodnego z hipotezą |
|---|---|---|
| Równomierność rozkładu łącznego | $[\cdot]$ | — |
| Warunkowa równomierność rozkładu:<br>zmiennej $Y$ względem $X$<br>zmiennej $X$ względem $Y$ | <br>$[X]$<br>$[Y]$ | <br>$\{X\}$<br>$\{Y\}$ |
| Niezależność stochastyczna | $[X][Y]$ | $\{X\}$ oraz $\{Y\}$ |
| Dowolny rozkład zmiennych | $[XY]$ | $\{XY\}$ |

Wprowadzane oznaczenia hipotez korespondują z rozkładami brzegowymi koniecznymi do jednoznacznego określenia rozkładu łącznego zgodnego z daną hipotezą. Konwencji tej będziemy przestrzegać w dalszej części pracy. Oznaczenie $[XY]$ sugeruje że pomiędzy zmiennymi zachodzi pewien związek (a dokładniej, nie muszą być niezależne), dlatego znajdują się w jednym nawiasie kwadratowym.

### 1.1.2 Hipotezy dla trzech zmiennych nominalnych

Do tej pory zostały przedstawione hipotezy dotyczące rozkładu dwóch zmiennych nominalnych. Jednak nawet w sytuacji gdy w centrum zainteresowania badacza jest związek pomiędzy dwiema zmiennymi, ważna jest odpowiedź na pytanie czy istnieją

jakieś inne czynniki wpływające na siłę tego związku. Dla przykładu badając zależność pomiędzy wykształceniem a sytuacją zawodową, można zapytać, czy związek ten jest taki sam wśród kobiet i mężczyzn, w mieście i na wsi itd.... Może się zdarzyć, że badając całą zbiorowość ustalimy silny związek pomiędzy dwiema zmiennymi, podczas gdy w pozdbiorowościach wyróżnionych ze względu na trzecią zmienną, badane zmienne są niezależne. Bądź też na odwrót: zmienne są niezależne w całej zbiorowości podczas gdy w podzbiorowościach zmienne te są silnie ze sobą związane.

Możemy mieć więc do czynienia z pozorną zależnością bądź pozorną niezależnością zmiennych w całej zbiorowości. Z tego względu istotne jest uwzględnienie większej liczby zmiennych w przeprowadzanych analizach. Oczywiście wybór czynników, których wpływ chcemy kontrolować nie powinien być przypadkowy. Pomocna może okazać się znajomość badanego zjawiska, pewne przesłanki teoretyczne.

W tej części sformułowane zostaną hipotezy opisujące związki pomiędzy trzema zmiennymi, które mogą być mierzone na skali nominalnej. Większość z tych hipotez można sformułować za pomocą pojęć, które wprowadzone zostały przy omawianiu hipotez dla dwóch zmiennych tj. równomierności rozkładu, niezależności stochastycznej, bądź pojęć będących ich rozszerzeniem tj. warunkowej równomierności czy też warunkowej niezależności stochastycznej. Dla sformułowania jednej z hipotez konieczne będzie jednak wprowadzenie nowego pojęcia, tj. stosunku szans.

Punktem wyjścia jest więc rozkład łączny trzech zmiennych. Rozkład taki przedstawiony został w Tabeli 1.8. W porównaniu do tabeli 1.1 uwzględniona została trzecia zmienna $Z$, która przyjmuje wartości $z_1$, $z_2$, $\ldots z_k \ldots z_t$, gdzie $t$ jest liczbą kategorii zmiennej $Z$. Uwzględnione zostały również rozkłady brzegowe zmiennych. Do tej pory mówiliśmy jedynie o rozkładach brzegowych uwzględniających wartości jednej zmiennej, możliwe jest jednak wyodrębnienie rozkładów uwzględniających większą liczbę zmiennych. Przykładowo, sumując prawdopodobieństwa $\pi_{ijk}^{XYZ}$, po indeksie zmiennej $Z$ otrzymujemy rozkład brzegowy zmiennych $X$ oraz $Y$ określający prawdopodobieństwa $\pi_{ij}^{XY}$. Rozkład ten oznaczać będziemy $\{XY\}$. Analogicznie uzyskać można rozkłady brzegowe $\{XZ\}$, $\{YZ\}$. W tabeli uwzględnione są również rozkłady brzegowe powstałe przez zsumowanie po indeksach dwóch zmiennych tj. $\{X\}$, $\{Y\}$, oraz $\{Z\}$.

Rozkład brzegowy uwzględniający większą liczbę zmiennych np. $\{XY\}$ można traktować jako rozkład *zmiennej złożonej*, uwzględniającej wszystkie $r \cdot c$ kombinacje obydwu zmiennych. Podobnie można mówić o warunkowym rozkładzie zmiennej złożonej np. rozkładzie zmiennej $XY$ w podzbiorowości wyróżnionej ze względu na zmienną $Z$. Pojęcie zmiennej złożonej jest wygodne do opisu rozkładu łącznego trzech

Tabela 1.8: Rozkład łączny trzech zmiennych $X$, $Y$ i $Z$

| | $Z = z_1$ | | | | $Z = z_2$ | | | | ... | $Z = z_t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $y_1$ | $y_2$ | ... | $y_c$ | ... | |
| $x_1$ | $\pi_{111}^{XYZ}$ | $\pi_{121}^{XYZ}$ | | | $\pi_{112}^{XYZ}$ | $\pi_{122}^{XYZ}$ | | | ... | |
| $x_2$ | $\pi_{211}^{XYZ}$ | $\pi_{221}^{XYZ}$ | | | $\pi_{212}^{XYZ}$ | $\pi_{222}^{XYZ}$ | | | ... | |
| $\vdots$ | | | | | | | | | | |
| $x_r$ | | | ... | $\pi_{rc1}^{XYZ}$ | | | ... | $\pi_{rc2}^{XYZ}$ | ... | |

| $\{XY\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $\{X\}$ |
| $x_1$ | $\pi_{11}^{XY}$ | $\pi_{12}^{XY}$ | | | $\pi_1^X$ |
| $x_2$ | $\pi_{21}^{XY}$ | $\pi_{22}^{XY}$ | | | $\pi_2^X$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\pi_{rc}^{XY}$ | $\pi_r^X$ |
| $\{Y\}$ | $\pi_1^Y$ | $\pi_2^Y$ | ... | $\pi_c^Y$ | 1 |

| $\{XZ\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{X\}$ |
| $x_1$ | $\pi_{11}^{XZ}$ | $\pi_{12}^{XZ}$ | | | $\pi_1^X$ |
| $x_2$ | $\pi_{21}^{XZ}$ | $\pi_{22}^{XZ}$ | | | $\pi_2^X$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\pi_{rt}^{XZ}$ | $\pi_r^X$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

| $\{YZ\}$ | | | | | |
|---|---|---|---|---|---|
| $Y \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{Y\}$ |
| $y_1$ | $\pi_{11}^{YZ}$ | $\pi_{12}^{YZ}$ | | | $\pi_1^Y$ |
| $y_2$ | $\pi_{21}^{YZ}$ | $\pi_{22}^{YZ}$ | | | $\pi_2^Y$ |
| $\vdots$ | | | | | |
| $y_c$ | | | ... | $\pi_{ct}^{YZ}$ | $\pi_c^Y$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

i większej liczby zmiennych, dlatego będzie ono wykorzystywane w dalszej części pracy.

Rozkład przedstawiony w tabeli 1.8 nie zakłada niczego odnośnie rozkładu łącznego trzech zmiennych. Tak jak w przypadku dwóch zmiennych hipotezy przedstawiać będą rozkłady o strukturze prostszej. Prezentację rozpocznie najprostsza sytuacja równomierności rozkładu łącznego a w dalszej kolejności prezentowane będą hipotezy bardziej skomplikowane. Poszczególne hipotezy będą również formułowane w odniesieniu do trzech przykładowych zmiennych. Będą to — tak jak poprzednio — udział w wyborach (dwuwartościowa zmienna $X$), wykształcenie (trójwartościowa zmienna $Y$) i dodatkowo uwzględniona zostanie informacja o płci (zmienna $Z$).

**Hipoteza $[\cdot]$ o równomierności rozkładu łącznego trzech zmiennych**

W rozkładzie zgodnym z tą hipotezą wszystkie prawdopodobieństwa są takie same. Zgodnie z tą hipotezą każde prawdopodobieństwo rozkładu łącznego można zapisać jako:

$$\pi_{ijk}^{XYZ} = \frac{1}{r \cdot c \cdot t} \qquad \text{dla wszystkich kombinacji } i, j, k. \tag{1.10}$$

Tabela 1.9: Rozkład łączny zgodny z hipotezą równomierności

| | $Z = z_1$ | | | | $Z = z_2$ | | | | $\dots$ | $Z = z_t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | $\dots$ | $y_c$ | $y_1$ | $y_2$ | $\dots$ | $y_c$ | $\dots$ | |
| $x_1$ | $\frac{1}{rct}$ | $\frac{1}{rct}$ | | | $\frac{1}{rct}$ | $\frac{1}{rct}$ | | | $\dots$ | |
| $x_2$ | $\frac{1}{rct}$ | $\frac{1}{rct}$ | | | $\frac{1}{rct}$ | $\frac{1}{rct}$ | | | $\dots$ | |
| $\vdots$ | | | | | | | | | | |
| $x_r$ | | | $\dots$ | $\frac{1}{rct}$ | | | $\dots$ | $\frac{1}{rct}$ | $\dots$ | |

| $\{XY\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | $\dots$ | $y_c$ | $\{X\}$ |
| $x_1$ | $\frac{1}{rc}$ | $\frac{1}{rc}$ | | | $\frac{1}{r}$ |
| $x_2$ | $\frac{1}{rc}$ | $\frac{1}{rc}$ | | | $\frac{1}{r}$ |
| $\vdots$ | | | | | |
| $x_r$ | | | $\dots$ | $\frac{1}{rc}$ | $\frac{1}{r}$ |
| $\{Y\}$ | $\frac{1}{c}$ | $\frac{1}{c}$ | $\dots$ | $\frac{1}{c}$ | 1 |

| $\{XZ\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Z$ | $z_1$ | $z_2$ | $\dots$ | $z_t$ | $\{X\}$ |
| $x_1$ | $\frac{1}{rt}$ | $\frac{1}{rt}$ | | | $\frac{1}{r}$ |
| $x_2$ | $\frac{1}{rt}$ | $\frac{1}{rt}$ | | | $\frac{1}{r}$ |
| $\vdots$ | | | | | |
| $x_r$ | | | $\dots$ | $\frac{1}{rt}$ | $\frac{1}{r}$ |
| $\{Z\}$ | $\frac{1}{t}$ | $\frac{1}{t}$ | $\dots$ | $\frac{1}{t}$ | 1 |

| $\{YZ\}$ | | | | | |
|---|---|---|---|---|---|
| $Y \backslash Z$ | $z_1$ | $z_2$ | $\dots$ | $z_t$ | $\{Y\}$ |
| $y_1$ | $\frac{1}{ct}$ | $\frac{1}{ct}$ | | | $\frac{1}{c}$ |
| $y_2$ | $\frac{1}{ct}$ | $\frac{1}{ct}$ | | | $\frac{1}{c}$ |
| $\vdots$ | | | | | |
| $y_c$ | | | $\dots$ | $\frac{1}{ct}$ | $\frac{1}{c}$ |
| $\{Z\}$ | $\frac{1}{t}$ | $\frac{1}{t}$ | $\dots$ | $\frac{1}{t}$ | 1 |

Tabela 1.9 stanowi ilustrację tej hipotezy. Do wyznaczenia rozkładu zgodnego z tą hipotezą wystarczy informacja o liczbie kategorii wszystkich zmiennych tj. $r$, $c$, $t$. Podobnie wyznaczamy rozkłady brzegowe, tj. do wyznaczenia rozkładu brzegowego $\{XZ\}$ wystarcza informacja o liczbie kategorii obydwu zmiennych. Hipotezę tę bę-

dziemy oznaczać $[\cdot]$, ponieważ do wyznaczenia rozkładu nie są potrzebne informacje o rozkładach brzegowych.

Dla podanego przykładu hipoteza ta oznacza, że każda kombinacja trzech zmiennych — płci, wykształcenia i udziału w wyborach — jest tak samo prawdopodobna. Przykładowo kobiety o wykształceniu podstawowym, biorące udział w wyborach stanowią 1/12 całej zbiorowości, podobnie jak np. niegłosujący mężczyźni o wykształceniu średnim.

### Hipoteza $[Z]$ o warunkowej równomierności rozkładu łącznego dwóch zmiennych $X$ i $Y$ względem trzeciej zmiennej $Z$

Zgodnie z hipotezą tego typu rozkład łączny dwóch zmiennych np. $X$ oraz $Y$ jest równomierny w każdej podzbiorowości wyróżnionej przez trzecią zmienną, w tym przypadku zmienną $Z$. Mówiąc inaczej, jeśli potraktujemy zmienną $XY$ jako zmienną złożoną, to jej rozkład warunkowy względem zmiennej $Z$ jest równomierny. Formalnie:

$$\pi_{ij(k)}^{XY(Z)} = \frac{1}{r \cdot c} \qquad \text{dla wszystkich kombinacji } i, j, k. \tag{1.11}$$

Zgodnie z powyższym prawdopodobieństwo rozkładu łącznego jest równe:

$$\pi_{ijk}^{XYZ} = \frac{\pi_k^Z}{r \cdot c} \qquad \text{dla wszystkich kombinacji } i, j, k. \tag{1.12}$$

Jak widać, aby określić rozkład oczekiwany zgodny z tą hipotezą potrzebne są informacje na temat rozkładu brzegowego zmiennej $Z$, który nie musi być równomierny.

Tabela 1.10 stanowi ilustrację powyższej hipotezy. Warto zwrócić uwagę, na to, że nie tylko rozkłady warunkowe zmiennej złożonej $XY$, ale również rozkład brzegowy tej zmiennej jest równomierny i każde prawdopodobieństwo jest równe $1/rc$. Podobnie rozkłady obydwu zmiennych $X$ oraz $Y$ są równomierne: zarówno rozkłady warunkowe względem zmiennej Z jak i rozkłady brzegowe obydwu zmiennych tj.

$$\pi_{i(k)}^{X(Z)} = \pi_i^X = \frac{1}{r},$$

oraz

$$\pi_{j(k)}^{Y(Z)} = \pi_j^Y = \frac{1}{c}.$$

Ponieważ zmienna $Z$ jest niezależna stochastycznie od zmiennej złożonej $XY$ jak i od obydwu zmiennych $X$ oraz $Y$, rozkłady warunkowe zmiennej $Z$ są identyczne w każdej podzbiorowości wyróżnionej przez obydwie zmienne i dla każdej kombinacji tych zmiennych. Rozkład warunkowy zmiennej $Z$ jest taki sam jak rozkład brzegowy tej zmiennej. tj. $\pi_{k(ij)}^{Z(XY)} = \pi_{k(i)}^{Z(X)} = \pi_{k(j)}^{Z(Y)} = \pi_k^Z$.

Tabela 1.10: Rozkład zgodny z hipotezą o warunkowej równomierności rozkładu łącznego dwóch zmiennych $X$ i $Y$ względem $Z$

| | $Z = z_1$ | | | | $Z = z_2$ | | | | ... | $Z = z_t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $y_1$ | $y_2$ | ... | $y_c$ | ... | |
| $x_1$ | $\frac{\pi_1^Z}{rc}$ | $\frac{\pi_1^Z}{rc}$ | | | $\frac{\pi_2^Z}{rc}$ | $\frac{\pi_2^Z}{rc}$ | | | ... | |
| $x_2$ | $\frac{\pi_1^Z}{rc}$ | $\frac{\pi_1^Z}{rc}$ | | | $\frac{\pi_2^Z}{rc}$ | $\frac{\pi_2^Z}{rc}$ | | | ... | |
| $\vdots$ | | | | | | | | | | |
| $x_r$ | | | ... | $\frac{\pi_1^Z}{rc}$ | | | ... | $\frac{\pi_2^Z}{rc}$ | ... | |

| $\{XY\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $\{X\}$ |
| $x_1$ | $\frac{1}{rc}$ | $\frac{1}{rc}$ | | | $\frac{1}{r}$ |
| $x_2$ | $\frac{1}{rc}$ | $\frac{1}{rc}$ | | | $\frac{1}{r}$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\frac{1}{rc}$ | $\frac{1}{r}$ |
| $\{Y\}$ | $\frac{1}{c}$ | $\frac{1}{c}$ | ... | $\frac{1}{c}$ | 1 |

| $\{XZ\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{X\}$ |
| $x_1$ | $\frac{\pi_1^Z}{r}$ | $\frac{\pi_2^Z}{r}$ | | | $\frac{1}{r}$ |
| $x_2$ | $\frac{\pi_1^Z}{r}$ | $\frac{\pi_2^Z}{r}$ | | | $\frac{1}{r}$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\frac{\pi_t^Z}{r}$ | $\frac{1}{r}$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

| $\{YZ\}$ | | | | | |
|---|---|---|---|---|---|
| $Y \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{Y\}$ |
| $y_1$ | $\frac{\pi_1^Z}{c}$ | $\frac{\pi_2^Z}{c}$ | | | $\frac{1}{c}$ |
| $y_2$ | $\frac{\pi_1^Z}{c}$ | $\frac{\pi_2^Z}{c}$ | | | $\frac{1}{c}$ |
| $\vdots$ | | | | | |
| $y_c$ | | | ... | $\frac{\pi_t^Z}{c}$ | $\frac{1}{c}$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

Hipoteza powyższa dla naszego przykładu, wskazuje, że każda kombinacja wykształcenia i udziału w wyborach jest tak samo prawdopodobna jeśli rozpatrujemy podzbiorowości kobiet i mężczyzn. Przykładowo, osoby o wykształceniu podstawowym, biorące udział w wyborach stanowią 1/6 kobiet, podobnie osoby o wykształceniu średnim, nie biorące udziału w wyborach stanowią 1/6 mężczyzn. Hipoteza nie głosi nic odnośnie odsetka kobiet i mężczyzn.

Powyższą hipotezę będziemy oznaczać $[Z]$. Analogicznie można sformułować jeszcze dwie hipotezy tego typu: $[X]$ oraz $[Y]$. Pierwsza z nich mówi o równomierności rozkładu łącznego zmiennych $Y$ i $Z$ względem $X$ (tj. każda kombinacja wykształcenia

i płci jest tak samo prawdopodobna w dwóch podzbiorowościach: osób biorących i nie biorących udziału w wyborach) a druga o równomierności rozkładu łącznego $X$ oraz $Z$ względem $Y$ (tj. każda kombinacja zmiennej opisującej udział w wyborach i płci jest tak samo prawdopodobna w trzech podzbiorowościach: osób o wykształceniu podstawowym, średnim lub wyższym).

### Hipoteza $[Y][Z]$ o warunkowej równomierności rozkładu zmiennej $X$ względem kombinacji dwóch niezależnych stochastycznie zmiennych $Y$ i $Z$

Kolejną hipotezę można przedstawić jako połączenie dwóch hipotez. Pierwsza z nich głosi, że zmienna $X$ ma rozkład równomierny w każdej pozbiorowości wyróżnionej przez złożoną zmienną $YZ$, tj. dla ustalonej kombinacji wartości zmiennych $Y$ i $Z$ wszystkie wartości zmiennej $X$ są jednakowo prawdopodobne. Druga z hipotez głosi, że zmienne tworzące zmienną złożoną tj. $Y$ oraz $Z$ są niezależne stochastycznie. Formalnie powyższą hipotezę można przedstawić jako:

$$\pi_{i(jk)}^{X(YZ)} = \frac{1}{r} \qquad \text{dla wszystkich kombinacji } i, j, k. \tag{1.13}$$

oraz

$$\pi_{jk} = \pi_j^Y \cdot \pi_k^Z \qquad \text{dla wszystkich kombinacji } i, j, k. \tag{1.14}$$

Z warunków 1.13 i 1.14 wynika, że prawdopodobieństwo rozkładu łącznego zgodnego z powyższą hipotezą można określić jako:

$$\pi_{ijk}^{XYZ} = \frac{\pi_j^Y \cdot \pi_k^Z}{r} \qquad \text{dla wszystkich kombinacji } i, j, k. \tag{1.15}$$

Rozkłady brzegowe potrzebne do jednoznacznego określenia powyższej hipotezy to $\{Y\}$ oraz $\{Z\}$.

Tabela 1.11 stanowi ilustrację rozkładu zgodnego z tą hipotezą. Brzegowy rozkład zmiennej $X$ jest równomierny. Podobnie równomierne są warunkowe rozkłady tej zmiennej względem zmiennych $Y$ i $Z$ tj. $\pi_{i(j)}^{X(Y)}$ oraz $\pi_{i(k)}^{X(Z)}$. W konsekwencji zmienna $X$ jest niezależna stochastycznie względem zmiennej $Y$, zmiennej $Z$ jak i zmiennej złożonej $YZ$. Zmienne $Y$ oraz $Z$ są niezależne stochastycznie nie tylko w całej zbiorowości, ale również w każdej podzbiorowości wyróżnionej ze względu na wartości zmiennej $X$.

Omawiana hipoteza dla trzech zmiennych z omawianego przykładu głosi, że dla każdej kombinacji wykształcenia i płci połowa osób brała udział w wyborach. Przykładowo, głosowała połowa kobiet o wykształceniu podstawowym, podobnie połowa mężczyzn o wykształceniu średnim, itd. Hipoteza nie głosi nic odnośnie rozkładu zmiennej opisującej płeć ani rozkładu wykształcenia, ale — co istotne — zmienne te

Tabela 1.11: Rozkład łączny zgodny z hipotezą o warunkowej równomierności zmiennej $X$ względem kombinacji dwóch niezależnych stochastycznie zmiennych $Y$ i $Z$

| | $Z = z_1$ | | | | $Z = z_2$ | | | | ... | $Z = z_t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $y_1$ | $y_2$ | ... | $y_c$ | ... | |
| $x_1$ | $\frac{\pi_1^Y \pi_1^Z}{r}$ | $\frac{\pi_2^Y \pi_1^Z}{r}$ | | | $\frac{\pi_1^Y \pi_2^Z}{r}$ | $\frac{\pi_2^Y \pi_2^Z}{r}$ | | | ... | |
| $x_2$ | $\frac{\pi_1^Y \pi_1^Z}{r}$ | $\frac{\pi_2^Y \pi_1^Z}{r}$ | | | $\frac{\pi_1^Y \pi_2^Z}{r}$ | $\frac{\pi_2^Y \pi_2^Z}{r}$ | | | ... | |
| $\vdots$ | | | | | | | | | | |
| $x_r$ | | | ... | $\frac{\pi_c^Y \pi_1^Z}{r}$ | | | ... | $\frac{\pi_c^Y \pi_2^Z}{r}$ | ... | |

| $\{XY\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $\{X\}$ |
| $x_1$ | $\frac{\pi_1^Y}{r}$ | $\frac{\pi_2^Y}{r}$ | | | $\frac{1}{r}$ |
| $x_2$ | $\frac{\pi_1^Y}{r}$ | $\frac{\pi_2^Y}{r}$ | | | $\frac{1}{r}$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\frac{\pi_c^Y}{r}$ | $\frac{1}{r}$ |
| $\{Y\}$ | $\pi_1^Y$ | $\pi_2^Y$ | ... | $\pi_c^Y$ | 1 |

| $\{XZ\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{X\}$ |
| $x_1$ | $\frac{\pi_1^Z}{r}$ | $\frac{\pi_2^Z}{r}$ | | | $\frac{1}{r}$ |
| $x_2$ | $\frac{\pi_1^Z}{r}$ | $\frac{\pi_2^Z}{r}$ | | | $\frac{1}{r}$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\frac{\pi_t^Z}{r}$ | $\frac{1}{r}$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

| $\{YZ\}$ | | | | | |
|---|---|---|---|---|---|
| $Y \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{Y\}$ |
| $y_1$ | $\pi_1^Y \pi_1^Z$ | $\pi_1^Y \pi_2^Z$ | | | $\pi_1^Y$ |
| $y_2$ | $\pi_2^Y \pi_1^Z$ | $\pi_2^Y \pi_2^Z$ | | | $\pi_2^Y$ |
| $\vdots$ | | | | | |
| $y_c$ | | | ... | $\pi_c^Y \pi_t^Z$ | $\pi_c^Y$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

są niezależne stochastycznie, zarówno wśród osób głosujących jak też nie biorących udziału w wyborach.

Powyższą hipotezę będziemy oznaczać $[Z][Y]$. Analogicznie można sformułować jeszcze dwie hipotezy tego typu: $[X][Z]$ oraz $[X][Y]$.

### Hipoteza $[X][Y][Z]$ o niezależności stochastycznej trzech zmiennych

W hipotezie tej nie zakładamy niczego odnośnie równomierności rozkładu którejkolwiek zmiennej czy też zmiennej złożonej. Podobnie jak w hipotezach poprzednich zakładamy jednak, że wszystkie trzy zmienne są od siebie niezależne stochastycznie. W tym miejscu omówione zostaną dokładniej cechy rozkładu łącznego przy niezależności trzech zmiennych. Jeśli trzy zmienne $X$, $Y$ oraz $Z$ są niezależne stochastycznie to:

- rozkłady warunkowe każdej zmiennej są takie same w podzbiorowościach wyróżnionych ze względu na kombinacje wartości pozostałych zmiennych np. warunkowe rozkłady zmiennej $X$ są takie same w każdej podzbiorowości wyróżnionej przez wartości zmiennej złożonej $YZ$
- Każda para zmiennych jest również niezależna stochastycznie, np. niezależne stochastycznie są zmienne $Y$ i $Z$.

Formalnie możemy to zapisać jako:

$$\pi_{i(jk)}^{X(YZ)} = \pi_i^X \qquad \text{dla wszystkich kombinacji } i,\, j,\, k \tag{1.16}$$

oraz:

$$\pi_{jk}^{YZ} = \pi_j^Y \cdot \pi_k^Z \qquad \text{dla wszystkich kombinacji } j,\, k. \tag{1.17}$$

To samo można powiedzieć o rozkładach warunkowych zmiennej $Y$ i zmiennej $Z$ względem pozostałych zmiennych. Połączenie powyższych warunków pozwala sformułować warunek analogiczny do 1.9. Tj.

$$\pi_{ijk}^{XYZ} = \pi_i^X \cdot \pi_j^Y \cdot \pi_k^Z \qquad \text{dla wszystkich kombinacji } i,\, j,\, k. \tag{1.18}$$

Prawdopodobieństwo rozkładu łącznego jest więc iloczynem odpowiednich prawdopodobieństw brzegowych. Informacje o rozkładach brzegowych $\{X\}$, $\{Y\}$, $\{Z\}$ wystarczają więc do jednoznacznego określenia rozkładu łącznego zmiennych.

Z tabeli 1.12 będącej ilustracją dla powyższej hipotezy widać również, że rozkłady warunkowe jednej zmiennej względem innej zmiennej np. $X$ względem $Y$ również są identyczne i równe brzegowemu rozkładowi tej zmiennej. Podobnie identyczne są warunkowe rozkłady dwóch zmiennych np. $X$ i $Y$ względem trzeciej zmiennej $Z$. Mówiąc inaczej zmienna złożona $XY$ jest niezależna stochastycznie od zmiennej $Z$.

Omawiana hipoteza dla trzech zmiennych z omawianego przykładu głosi, że dla każdej z sześciu kombinacji wykształcenia i płci głosował taki sam odsetek osób. Przykładowo, taki sam odsetek osób głosował wśród kobiet o wykształceniu podstawowym,

Tabela 1.12: Rozkład łączny zgodny z hipotezą o niezależności stochastycznej trzech zmiennych $X$, $Y$ i $Z$

| | $Z = z_1$ | | | | $Z = z_2$ | | | | ... | $Z = z_t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $X\backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $y_1$ | $y_2$ | ... | $y_c$ | ... | |
| $x_1$ | $\pi_1^X \pi_1^Y \pi_1^Z$ | $\pi_1^X \pi_2^Y \pi_1^Z$ | | | $\pi_1^X \pi_1^Y \pi_2^Z$ | $\pi_1^X \pi_2^Y \pi_2^Z$ | | | ... | |
| $x_2$ | $\pi_2^X \pi_1^Y \pi_1^Z$ | $\pi_2^X \pi_2^Y \pi_1^Z$ | | | $\pi_2^X \pi_1^Y \pi_2^Z$ | $\pi_2^X \pi_2^Y \pi_2^Z$ | | | ... | |
| $\vdots$ | | | | | | | | | | |
| $x_r$ | | | ... | | | | ... | | ... | |

| $\{XY\}$ | | | | | |
|---|---|---|---|---|---|
| $X\backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $\{X\}$ |
| $x_1$ | $\pi_1^X \pi_1^Y$ | $\pi_1^X \pi_2^Y$ | | | $\pi_1^X$ |
| $x_2$ | $\pi_2^X \pi_1^Y$ | $\pi_2^X \pi_2^Y$ | | | $\pi_2^X$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\pi_r^X \pi_c^Y$ | $\pi_r^X$ |
| $\{Y\}$ | $\pi_1^Y$ | $\pi_2^Y$ | ... | $\pi_c^Y$ | 1 |

| $\{XZ\}$ | | | | | |
|---|---|---|---|---|---|
| $X\backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{X\}$ |
| $x_1$ | $\pi_1^X \pi_1^Z$ | $\pi_1^X \pi_2^Z$ | | | $\pi_1^X$ |
| $x_1$ | $\pi_2^X \pi_1^Z$ | $\pi_2^X \pi_2^Z$ | | | $\pi_2^X$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\pi_r^X \pi_t^Z$ | $\pi_r^X$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

| $\{YZ\}$ | | | | | |
|---|---|---|---|---|---|
| $Y\backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{Y\}$ |
| $y_1$ | $\pi_1^Y \pi_1^Z$ | $\pi_1^Y \pi_2^Z$ | | | $\pi_1^Y$ |
| $y_2$ | $\pi_2^Y \pi_1^Z$ | $\pi_2^Y \pi_2^Z$ | | | $\pi_2^Y$ |
| $\vdots$ | | | | | |
| $y_c$ | | | ... | $\pi_c^Y \pi_t^Z$ | $\pi_c^Y$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

jak wśród mężczyzn o wykształceniu średnim, itd. Niezależne są też wykształcenie i płeć zarówno wśród osób głosujących jak też nie biorących udziału w wyborach. Hipoteza ta jest symetryczna, wynika z tego m. in. że rozkład wykształcenia jest identyczny w każdej podzbiorowosci wyróżnionej ze względu na udział w wyborach i płeć (przy czym te dwie zmienne są również niezależne ze względu na wykształcenie): przykładowo odsetek osób z wykształcaniem podstawowym jest taki sam wśród ko-

biet głosujących, jak wśród niegłosujących mężczyzn. Hipoteza nie głosi nic odnośnie rozkładu żadnej z wymienionych zmiennych.

### Hipoteza $[YZ]$ o warunkowej równomierności zmiennej X względem kombinacji wartości dwóch pozostałych zmiennych $Y$ i $Z$

Hipoteza ta głosi, że rozkład jednej zmiennej np. $X$ jest równomierny względem kombinacji wartości dwóch pozostałych zmiennych tj. $Y$ i $Z$. W przeciwieństwie do prezentowanej wcześniej hipotezy $[Y][Z]$ nie zakładamy jednak, że zmienne $Y$ i $Z$ są niezależne stochastycznie. Formalnie:

$$\pi_{i(jk)}^{X(YZ)} = \frac{1}{r} \qquad \text{dla wszystkich kombinacji } i,\ j,\ k. \tag{1.19}$$

Prawdopodobieństwo rozkładu łącznego zgodnego z powyższą hipotezą można określić jako:

$$\pi_{ijk}^{XYZ} = \frac{\pi_{jk}^{YZ}}{r} \qquad \text{dla wszystkich kombinacji } i,\ j,\ k. \tag{1.20}$$

Do określenia rozkładu łącznego związanego z powyższą hipotezą jest więc potrzebny rozkład brzegowy zmiennej dwuwymiarowej $\{YZ\}$ tj. informacje o prawdopodobieństwach $\pi_{jk}^{YZ}$. Ponieważ zmienne $Y$ i $Z$ mogą być zależne, rozkładu tego nie da się juz zredukować do iloczynu odpowiednich prawdopodobieństw brzegowych.

Tabela 1.13 przedstawia rozkład zgodny z powyższą hipotezą. Jak widać rozkład zmiennej $X$ jest równomierny również w podzbiowościach wyróżnionych względem zmiennej $Y$ i zmiennej $X$. Również brzegowy rozkład tej zmiennej jest równomierny tj.

$$\pi_{i(jk)}^{X(YZ)} = \pi_{i(j)}^{X(Y)} = \pi_{i(k)}^{X(Z)} = \pi_i^X = \frac{1}{r}.$$

Podobnie można zauważyć, że rozkład zmiennej dwuwartościowej $YZ$ jest taki sam w każdej podzbiorowości wyodrębnionej ze względu na wartości zmiennej $X$. Zmienne $YZ$ oraz $X$ są więc niezależne stochastycznie.

Nie można powiedzieć, czy hipoteza ta jest prostsza czy tez bardziej skomplikowana od hipotezy o niezależności stochastycznej trzech zmiennych. Z jednej strony dopuszczamy zależność między dwiema zmiennymi, z drugiej strony zakładamy, że rozkład trzeciej zmiennej jest równomierny. Jest to jednak hipoteza bardziej skomplikowana od hipotez $[Z][Y]$ i hipotez od niej prostszych.

Odnosząc tę hipotezę do wybranych trzech zmiennych, można powiedzieć, że dla każdej kombinacji wykształcenia i płci połowa osób głosowała. Przykładowo, głosowała połowa kobiet o wykształceniu podstawowym, podobnie połowa mężczyzn o wykształceniu średnim, itd. Hipoteza nie głosi nic o rozkładzie zmiennej opisującej płeć ani o rozkładzie wykształcenia, ponadto, zmienne te mogą być zależne.

Tabela 1.13: Rozkład łączny zgodny z hipotezą o warunkowej równomierności zmiennej $X$ względem kombinacji wartości dwóch zmiennych $Y$ i $Z$

| | $Z = z_1$ | | | | $Z = z_2$ | | | | ... | $Z = z_t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $y_1$ | $y_2$ | ... | $y_c$ | ... | |
| $x_1$ | $\frac{\pi_{11}^{YZ}}{r}$ | $\frac{\pi_{21}^{YZ}}{r}$ | | | $\frac{\pi_{12}^{YZ}}{r}$ | $\frac{\pi_{22}^{YZ}}{r}$ | | | ... | |
| $x_2$ | $\frac{\pi_{11}^{YZ}}{r}$ | $\frac{\pi_{21}^{YZ}}{r}$ | | | $\frac{\pi_{12}^{YZ}}{r}$ | $\frac{\pi_{22}^{YZ}}{r}$ | | | ... | |
| $\vdots$ | | | | | | | | | | |
| $x_r$ | | | ... | $\frac{\pi_{c1}^{YZ}}{r}$ | | | ... | $\frac{\pi_{c2}^{YZ}}{r}$ | ... | |

| $\{XY\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $\{X\}$ |
| $x_1$ | $\frac{\pi_1^Y}{r}$ | $\frac{\pi_2^Y}{r}$ | | | $\frac{1}{r}$ |
| $x_2$ | $\frac{\pi_1^Y}{r}$ | $\frac{\pi_2^Y}{r}$ | | | $\frac{1}{r}$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\frac{\pi_c^Y}{r}$ | $\frac{1}{r}$ |
| $\{Y\}$ | $\pi_1^Y$ | $\pi_2^Y$ | ... | $\pi_c^Y$ | 1 |

| $\{XZ\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{X\}$ |
| $x_1$ | $\frac{\pi_1^Z}{r}$ | $\frac{\pi_2^Z}{r}$ | | | $\frac{1}{r}$ |
| $x_2$ | $\frac{\pi_1^Z}{r}$ | $\frac{\pi_2^Z}{r}$ | | | $\frac{1}{r}$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\frac{\pi_t^Z}{r}$ | $\frac{1}{r}$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

| $\{YZ\}$ | | | | | |
|---|---|---|---|---|---|
| $Y \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{Y\}$ |
| $y_1$ | $\pi_{11}^{YZ}$ | $\pi_{12}^{YZ}$ | | | $\pi_1^Y$ |
| $y_2$ | $\pi_{21}^{YZ}$ | $\pi_{22}^{YZ}$ | | | $\pi_2^Y$ |
| $\vdots$ | | | | | |
| $y_c$ | | | ... | $\pi_{ct}^{YZ}$ | $\pi_c^Y$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

Analogicznie można sformułować dwie inne hipotezy tego typu: o równomierności rozkładu warunkowego zmiennej $Y$ względem zmiennej $XZ$, tj. $[XZ]$ oraz o równomierności rozkładu warunkowego zmiennej Z względem zmiennej $XY$, tj. $[XY]$. Zgodnie z pierwszą z nich dla każdej kombinacji udziału w wyborach i płci, 1/3 stanowią osoby o wykształceniu podstawowym, 1/3 – o wykształceniu średnim i 1/3 stanowią osoby posiadające wyższe wykształcenie. Natomiast zgodnie z hipotezą $[XY]$ rozkład płci jest równomierny, w każdej z sześciu podzbiorowości wyróżnionej ze względu

na wykształcenie i udział w wyborach, tj. wśród osób głosujących o wykształcaniu podstawowym połowę stanowią kobiety, podobnie wśród osób niegłosujących o wykształcaniu średnim, itd.

### Hipoteza $[X][YZ]$ o identyczności rozkładu zmiennej $X$ względem kombinacji dwóch pozostałych zmiennych $Y$ i $Z$

Ten typ hipotezy głosi, że rozkład jednej zmiennej np. $X$ jest taki sam w każdej podzbiorowości wyróżnionej przez kombinacje zmiennych $Y$ i $Z$. Mówiąc inaczej zmienna $X$ i zmienna $YZ$ są niezależne stochastycznie. Hipoteza ta jest bardziej skomplikowana od poprzednio prezentowanych hipotez. Nie zakładamy nic na temat równomierności rozkładu którejkolwiek ze zmiennych (tak jak zakładaliśmy w hipotezie $[YZ]$). Dopuszczamy również że jedna z par zmiennych może być zależna stochastycznie (w przeciwieństwie do hipotezy $[X][Z][Y]$ i prostszych od niej). Formalnie powyższą hipotezę można zapisać jako:

$$\pi^{X(YZ)}_{i(jk)} = \pi^{X}_{i} \qquad \text{dla wszystkich kombinacji } i,\ j,\ k. \tag{1.21}$$

Prawdopodobieństwo rozkładu łącznego zgodnego z powyższą hipotezą można określić jako:

$$\pi^{XYZ}_{ijk} = \pi^{X}_{i} \cdot \pi^{YZ}_{jk} \qquad \text{dla wszystkich kombinacji } i,\ j,\ k. \tag{1.22}$$

Do jednoznacznego określenia rozkładu tej zmiennej potrzebny jest rozkład $\{X\}$ oraz rozkład zmiennej dwuwymiarowej $\{YZ\}$. Tabela 1.14 stanowi ilustrację powyższej hipotezy.

Można zauważyć, że wszystkie rozkłady warunkowe zmiennej $X$ są identyczne i takie same jak rozkład tej zmiennej w całej zbiorowości tj:

$$\pi^{X(YZ)}_{i(jk)} = \pi^{X(Y)}_{i(j)} = \pi^{X(Z)}_{i(k)} = \pi^{X}_{i},$$

czyli poza tym, że zmienna $X$ jest niezależna względem zmiennej $YZ$, możemy również mówić o niezależności par $X$ i $Y$ oraz $X$ i $Z$. Pary te są niezależne zarówno w całej zbiorowości jak i podzbiorowościach wyróżnionych ze względu na trzecią zmienną, np. zmienne $X$ i $Y$ są niezależne w każdej podzbiorowości wyróżnionej przez zmienną $Z$.

Co głosi powyższa hipoteza dla trzech zmiennych z omawianego przykładu? Zgodnie z nią dla każdej kombinacji wykształcenia i płci głosował taki sam odsetek osób. Przykładowo, taki sam odsetek osób głosował wśród kobiet o wykształceniu podstawowym, jak wśród mężczyzn o wykształceniu średnim. Wykształcenie i płeć mogą być zależne od siebie, ponadto hipoteza nie głosi nic odnośnie rozkładu żadnej z wymienionych zmiennych.

Tabela 1.14: Rozkład łączny zgodny z hipotezą o identyczności rozkładu zmiennej $X$ względem kombinacji wartości zmiennych $Y$ i $Z$

| | $Z = z_1$ | | | | $Z = z_2$ | | | | ... | $Z = z_t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $y_1$ | $y_2$ | ... | $y_c$ | ... | |
| $x_1$ | $\pi_1^X \pi_{11}^{YZ}$ | $\pi_1^X \pi_{21}^{YZ}$ | | | $\pi_1^X \pi_{12}^{YZ}$ | $\pi_1^X \pi_{22}^{YZ}$ | | | ... | |
| $x_2$ | $\pi_2^X \pi_{11}^{YZ}$ | $\pi_2^X \pi_{21}^{YZ}$ | | | $\pi_2^X \pi_{12}^{YZ}$ | $\pi_2^X \pi_{22}^{YZ}$ | | | ... | |
| $\vdots$ | | | | | | | | | | |
| $x_r$ | | | ... | $\pi_r^X \pi_{c1}^{YZ}$ | | | ... | $\pi_r^X \pi_{c2}^{YZ}$ | ... | |

| $\{XY\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $\{X\}$ |
| $x_1$ | $\pi_1^X \pi_1^Y$ | $\pi_1^X \pi_2^Y$ | | | $\pi_1^X$ |
| $x_2$ | $\pi_2^X \pi_1^Y$ | $\pi_2^X \pi_2^Y$ | | | $\pi_2^X$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\pi_r^X \pi_c^Y$ | $\pi_r^X$ |
| $\{Y\}$ | $\pi_1^Y$ | $\pi_2^Y$ | ... | $\pi_c^Y$ | 1 |

| $\{XZ\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{X\}$ |
| $x_1$ | $\pi_1^X \pi_1^Z$ | $\pi_1^X \pi_2^Z$ | | | $\pi_1^X$ |
| $x_2$ | $\pi_2^X \pi_1^Z$ | $\pi_2^X \pi_2^Z$ | | | $\pi_2^X$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\pi_r^X \pi_t^Z$ | $\pi_r^X$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

| $\{YZ\}$ | | | | | |
|---|---|---|---|---|---|
| $Y \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{Y\}$ |
| $y_1$ | $\pi_{11}^{YZ}$ | $\pi_{12}^{YZ}$ | | | $\pi_1^Y$ |
| $y_2$ | $\pi_{21}^{YZ}$ | $\pi_{22}^{YZ}$ | | | $\pi_2^Y$ |
| $\vdots$ | | | | | |
| $y_c$ | | | ... | $\pi_{ct}^{YZ}$ | $\pi_c^Y$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

Analogicznie możliwe jest sformułowanie hipotez $[XZ][Y]$ i $[XY][Z]$. W pierwszym przypadku rozkład wykształcenia jest identyczny, dla każdej z sześciu podzbiorowości wyróżnianych ze względu na płeć i udział w wyborach: dla mężczyzn niegłosujących, dla głosujących kobiet itd. W drugim przypadku rozkład płci jest identyczny dla każdej z podzbiorowości, jakie można wyróżnić ze względu na wykształcenie i udział w wyborach.

### Hipoteza $[XZ][YZ]$ o warunkowej niezależności dwóch zmiennych $X$ i $Y$ względem trzeciej zmiennej $Z$

Hipoteza tego typu głosi, że dwie zmienne np. $X$ i $Y$ są niezależne stochastycznie w każdej podzbiorowości wyróżnionej przez trzecią zmienną $Z$. W takiej sytuacji:

$$\pi_{ij(k)}^{XY(Z)} = \pi_{i(k)}^{X(Z)} \cdot \pi_{j(k)}^{Y(Z)} \qquad \text{dla wszystkich kombinacji } i, j, k. \tag{1.23}$$

Jeśli obydwie strony równania 1.23 pomnożymy przez prawdopodobieństwo brzegowe zmiennej Z okaże się, że prawdopodobieństwo rozkładu łącznego jest równe:

$$\pi_{ijk}^{XYZ} = \pi_{ik}^{XZ} \cdot \pi_{j(k)}^{Y(Z)} = \frac{\pi_{ik}^{XZ} \cdot \pi_{jk}^{YZ}}{\pi_{k}^{Z}} \qquad \text{dla wszystkich kombinacji } i, j, k. \tag{1.24}$$

Do wyznaczenia rozkładu łącznego związanego z tą hipotezą potrzebne są więc informacje o rozkładach brzegowych dwóch zmiennych dwuwymiarowych $\{XZ\}$ oraz $\{YZ\}$. Tabela 1.15 stanowi ilustrację powyższej hipotezy.

Warto zwrócić uwagę na to, że mimo, że zmienne np. $X$ i $Y$ są niezależne stochastycznie w każdej podzbiorowości wyróżnionej przez trzecią zmienną $Z$, to w całej zbiorowości mogą być zależne stochastycznie. Wyrażenie:

$$\pi_{ij}^{XY} = \sum_{k=1}^{t} \frac{\pi_{ik}^{XZ} \pi_{jk}^{YZ}}{\pi_{k}^{Z}}$$

nie musi bowiem być równe iloczynowi $\pi_i^X \pi_j^Y$. Sytuacja jest więc inna niż w modelu $[X][YZ]$. Niezależność warunkowa zmiennych $X$ i $Y$ ze względu na zmienną $Z$ przekłada się na niezależność w całej zbiorowości, jeśli spełniony jest jeden lub obydwa warunki:

- Zmienne $X$ i $Z$ są niezależne warunkowo ze względu na zmienną $Y$.
- Zmienne $Y$ i $Z$ są niezależne warunkowo ze względu na zmienną $X$.

Wynika to z *twierdzenia o agregacji* (*collapsibility theorem*) [3], które w bardziej ogólnej formie przedstawione zostało w Aneksie. Hipoteza $[XZ][YZ]$ — w przeciwieństwie do hipotezy $[X][YZ]$ — nie zakłada żadnego z powyższych warunków, stąd zmienne $X$ i $Y$ nie muszą być zależne w całej zbiorowości.

Dla omawianego przykładu, powyższa hipoteza głosi, że wykształcenie i udział w wyborach są niezależne, jeśli rozpatrujemy każdą z podzbiorowości, jakie można wyróżnić ze względu na płeć. Przypuśćmy, że rozpatrujemy zbiorowość mężczyzn: wśród nich odsetek osób głosujących jest taki sam dla osób z wykształceniem podstawowym, średnim i wyższym. Gdybyśmy rozpatrywali wszystkie osoby nie musiałoby to

[3] Porównaj Bishop i inni (1975), str 47. Agresti, (1984), str. 37; Hagenaars, (1990), str. 68.

Tabela 1.15: Rozkład łączny zgodny z hipotezą o niezależności warunkowej zmiennych $X$ i $Y$ względem $Z$

| | $Z = z_1$ | | | | $Z = z_2$ | | | | ... | $Z = z_t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $y_1$ | $y_2$ | ... | $y_c$ | ... | |
| $x_1$ | $\frac{\pi_{11}^{XZ}\pi_{11}^{YZ}}{\pi_1^Z}$ | $\frac{\pi_{11}^{XZ}\pi_{21}^{YZ}}{\pi_1^Z}$ | | | $\frac{\pi_{12}^{XZ}\pi_{12}^{YZ}}{\pi_2^Z}$ | $\frac{\pi_{12}^{XZ}\pi_{22}^{YZ}}{\pi_2^Z}$ | | | ... | |
| $x_2$ | $\frac{\pi_{21}^{XZ}\pi_{11}^{YZ}}{\pi_1^Z}$ | $\frac{\pi_{21}^{XZ}\pi_{21}^{YZ}}{\pi_1^Z}$ | | | $\frac{\pi_{22}^{XZ}\pi_{12}^{YZ}}{\pi_2^Z}$ | $\frac{\pi_{22}^{XZ}\pi_{22}^{YZ}}{\pi_2^Z}$ | | | ... | |
| $\vdots$ | | | | | | | | | | |
| $x_r$ | | | ... | $\frac{\pi_{r1}^{XZ}\pi_{c1}^{YZ}}{\pi_1^Z}$ | | | ... | $\frac{\pi_{r2}^{XZ}\pi_{c2}^{YZ}}{\pi_2^Z}$ | ... | |

| $\{XY\}$ | | | | | |
|---|---|---|---|---|---|
| $Y \backslash Y$ | $y_1$ | $y_2$ | ... | $y_c$ | $\{X\}$ |
| $x_1$ | $\sum_{k=1}^{t} \frac{\pi_{1k}^{XZ}\pi_{1k}^{YZ}}{\pi_k^Z}$ | $\sum_{k=1}^{t} \frac{\pi_{1k}^{XZ}\pi_{2k}^{YZ}}{\pi_k^Z}$ | | | $\pi_1^X$ |
| $x_2$ | $\sum_{k=1}^{t} \frac{\pi_{2k}^{XZ}\pi_{1k}^{YZ}}{\pi_k^Z}$ | $\sum_{k=1}^{t} \frac{\pi_{2k}^{XZ}\pi_{2k}^{YZ}}{\pi_k^Z}$ | | | $\pi_2^X$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\sum_{k=1}^{t} \frac{\pi_{rk}^{XZ}\pi_{ck}^{YZ}}{\pi_k^Z}$ | $\pi_r^X$ |
| $\{Y\}$ | $\pi_1^Y$ | $\pi_2^Y$ | ... | $\pi_c^Y$ | 1 |

| $\{XZ\}$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{X\}$ |
| $x_1$ | $\pi_{11}^{XZ}$ | $\pi_{12}^{XZ}$ | | | $\pi_1^X$ |
| $x_2$ | $\pi_{21}^{XZ}$ | $\pi_{22}^{XZ}$ | | | $\pi_2^X$ |
| $\vdots$ | | | | | |
| $x_r$ | | | ... | $\pi_{rt}^{XZ}$ | $\pi_r^X$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

| $\{YZ\}$ | | | | | |
|---|---|---|---|---|---|
| $Y \backslash Z$ | $z_1$ | $z_2$ | ... | $z_t$ | $\{Y\}$ |
| $y_1$ | $\pi_{11}^{YZ}$ | $\pi_{12}^{YZ}$ | | | $\pi_1^Y$ |
| $y_2$ | $\pi_{21}^{YZ}$ | $\pi_{22}^{YZ}$ | | | $\pi_2^Y$ |
| $\vdots$ | | | | | |
| $y_c$ | | | ... | $\pi_{ct}^{YZ}$ | $\pi_c^Y$ |
| $\{Z\}$ | $\pi_1^Z$ | $\pi_2^Z$ | ... | $\pi_t^Z$ | 1 |

jednak być prawdą, przykładowo odsetek osób głosujących mógłby być inny wśród osób z wykształceniem podstawowym i wykształceniem średnim. Hipoteza nie głosi nic odnośnie rozkładu żadnej z wymienionych zmiennych. Ponadto, zarówno udział w wyborach, jak też wykształcenie mogą być zależne od płci.

Analogicznie można sformułować dwie hipotezy tego samego typu $[XY][XZ]$ oraz $[XY][YZ]$. W pierwszym przypadku, zarówno wśród osób głosujących, jak również

niegłosujących wykształcenie i płeć są niezależne. Hipoteza $[XY][YZ]$ głosi, że udział w wyborach i płeć są niezależne ze względu na wykształcenie. Przykładowo, wśród osób z wykształceniem podstawowym, odsetek osób głosujących jest taki sam wśród kobiet i wśród mężczyzn.

### Hipoteza $[XY][XZ][YZ]$ o równości warunkowych stosunków szans

Tak jak zostało zasygnalizowane wcześniej, nie wszystkie hipotezy dotyczące rozkładu trzech zmiennych, formułowane na gruncie modeli logarytmiczno—liniowych mają postać hipotez elementarnych, tj. nie wszystkie dają się przedstawić za pomocą pojęć warunkowej bądź bezwarunkowej niezależności stochastycznej i równomierności rozkładu. Do sformułowania kolejnej hipotezy pomocne jest pojęcie opisujące związek pomiędzy różnymi kategoriami dwóch zmiennych, mianowicie *stosunek szans*.

W modelach logarytmiczno–liniowych do opisu rozkładu zmiennej często wykorzystuje się pojęcie *szansy* (*odd*). Szansę, że zmienna $X$ przyjmuje raczej wartość $x_a$ aniżeli wartość $x_b$ oznaczać będziemy $\Omega^X_{a/b}$ i jest ona równa ilorazowi dwóch prawdopodobieństw związanych z tą zmienną tj.

$$\Omega^X_{a/b} = \frac{\pi^X_a}{\pi^X_b}. \tag{1.25}$$

Podobnie zdefiniować można szansę warunkową określającą, że zmienna $X$ przyjmie raczej wartość $a$ niż wartość $b$ w zbiorowości $Y = y_j$:

$$\Omega^{X\ (Y)}_{a/b(j)} = \frac{\pi^{XY}_{aj}}{\pi^{XY}_{bj}}. \tag{1.26}$$

W tabeli 1.16 mamy podany fikcyjny rozkład dotyczący uczestnictwa w wyborach i wykształcenia. W całej zbiorowości 77% osób głosowało, 23% osób nie brało udziału w wyborach. Szansa obliczona dla tych dwóch kategorii wynosi $\Omega^X_{1/2} = 3,34$ co oznacza, ze na jedną osobę, która nie głosowała przypada więcej niż 3 osoby głosujące. Należy zauważyć, ze proporcje te są różne w zależności od wykształcenia. Informują nas o tym warunkowe szanse. Są one równe $\Omega^{X\ (Y)}_{2/1(1)} = 2$ dla osób z wykształceniem podstawowym, $\Omega^{X\ (Y)}_{2/1(2)} = 2,87$ dla osób z wykształceniem średnim oraz $\Omega^{X\ (Y)}_{2/1(3)} = 9$ dla osób z wykształceniem wyższym.

*Stosunek szans* pozwala porównywać warunkowe szanse dla dwóch zbiorowości. Jeśli porównujemy szanse dla kategorii $x_a$ i kategorii $x_b$ zmiennej $X$ w dwóch zbiorowościach wyróżnionych ze względu na zmienną $Y$, tj. $y_c$, $y_d$, wówczas otrzymujemy:

$$\Theta^{X\ Y}_{a/b;c/d} = \frac{\Omega^{X\ (Y)}_{a/b(c)}}{\Omega^{X\ (Y)}_{a/b(d)}} = \frac{\pi^{XY}_{ac}/\pi^{XY}_{bc}}{\pi^{XY}_{ad}/\pi^{XY}_{bd}} = \frac{\pi^{XY}_{ac} \cdot \pi^{XY}_{bd}}{\pi^{XY}_{ad} \cdot \pi^{XY}_{bc}}. \tag{1.27}$$

Tabela 1.16: Wykształcenie a udział w wyborach (w procentach, dane fikcyjne)

| | Wykształcenie($Y$) | | | |
|---|---|---|---|---|
| Czy uczestniczył w wyborach? ($X$) | 1. Podstawowe | 2. Średnie | 3. Wyższe | Ogółem |
| 1. Nie | 5,0 | 15,5 | 2,5 | 23,0 |
| 2. Tak | 10,0 | 44,5 | 22,5 | 77,0 |
| Ogółem | 15,0 | 60,0 | 25,0 | 100,0 |

Dla przykładu z tabeli 1.16 stosunek szans wyróżniony dla dwóch kategorii opisujących uczestnictwo w wyborach oraz dwóch kategorii wykształcenia: podstawowego i średniego wynosi $\Theta^{X\ Y}_{2/1;2/1} = 2,87/2 = 1,43$. Oznacza to, że proporcja osób głosujących do osób niegłosujących wśród osób z wykształceniem średnim jest 1,5 razy większa niż wśród osób z wykształceniem podstawowym. Porównując szanse dotyczące udziału w wyborach między osobami z wykształceniem wyższym i średnim otrzymujemy $\Theta^{X\ Y}_{2/1;3/2} = 3,13$. Możemy więc stwierdzić, że różnice pomiędzy wykształceniem wyższym i średnim odnośnie udziału w wyborach są większe niż pomiędzy wykształceniem średnim i podstawowym. Możliwe jest również wyznaczenie stosunku szans dla skrajnych kategorii wykształcenia, tj. wykształcenia wyższego i podstawowego. Łatwo dostrzec, ze będzie on iloczynem dwóch powyższych stosunków szans i wynosi $\Theta^{X\ Y}_{2/1;3/2} = 4,5$. Jak widać, kategorie obydwu zmiennych są zależne. Im wyższe jest wykształcenie, tym większe prawdopodobieństwo głosowania.

Wzór 1.27 pokazuje, że stosunek szans jest wielkością symetryczną. Tak więc wielkość $\Theta^{X\ Y}_{2/1;3/1} = 4,5$ mówi nam również, że wśród osób głosujących proporcja pomiędzy osobami z wykształceniem wyższym w stosunku do osób z wykształceniem podstawowym jest 4,5 razy wyższa niż wśród nie uczestniczących w wyborach. Warto zwrócić uwagę na inną ważną własność stosunku szans. Jeśli liczebności w dowolnym wierszu lub w dowolnej kolumnie zostaną przemnożone przez stałą nie wpływa to na wielkość tego wyrażenia. Łatwo to dostrzec patrząc na formułę 1.27. Przypuśćmy, że dwukrotnie zwiększają się liczebności związane z kategorią $x_a$. Wpływa to w takim samym stopniu na prawdopodobieństwa $\pi^{XY}_{ac}$ oraz $\pi^{XY}_{ad}$, przy czym pierwsza z nich jest w liczniku a druga w mianowniku. Jest to pożądana własność, gdyż stosunki szans służą do opisu związku między zmiennymi i dlatego ich wielkość nie powinna zależeć od rozkładów brzegowych zmiennych.

Należy zauważyć, że gdyby obydwie zmienne były niezależne stochastycznie, wówczas wszystkie stosunki szans byłyby równe 1. Stosunki szans opisują więc zależność pomiędzy poszczególnymi kategoriami zmiennych. Jeżeli wartość stosunku szans jest

większa od 1 a kategorie obydwu zmiennych uporządkowane w pewien sposób, mamy do czynienia z zależnością pozytywną. Jeżeli wartość stosunku szans jest mniejsza od 1 wówczas można mówić o zależności negatywnej.

Tabela 1.17: Wykształcenie, uczestnictwo w wyborach a płeć — rozkład łączny (w procentach, dane fikcyjne)

| 1. Kobiety ($Z = 1$) | | | |
|---|---|---|---|
| Czy uczestniczył w wyborach? ($X$) | Wykształcenie ($Y$) | | |
| | 1. Podstawowe | 2. Średnie | 3. Wyższe |
| 1. Nie | 3,0 | 5,0 | 1,5 |
| 2. Tak | 7,0 | 20,0 | 13,5 |
| 2. Mężczyźni ($Z = 2$) | | | |
| Czy uczestniczył w wyborach ($X$) | Wykształcenie ($Y$) | | |
| | 1. Podstawowe | 2. Średnie | 3. Wyższe |
| 1. Nie | 2,0 | 10,5 | 1,0 |
| 2. Tak | 3,0 | 24,5 | 9,0 |

Aby porównywać siłę i kierunek związku pomiędzy kategoriami dwóch zmiennych w różnych podzbiorowościach można posługiwać się *warunkowymi stosunkami szans*. Dla omawianego przykładu interesujące może być porównanie, czy zależności pomiędzy wykształceniem a uczestnictwem w wyborach różnicuje płeć. Rozkład łączny trzech zmiennych przedstawia tabela 1.17. Porównując stosunki szans dla wykształcenia średniego i wyższego, jego wartość dla kobiet wynosi $\Theta^{X\ Y\ (Z)}_{2/1;3/2;(1)} = 3,85$ a dla mężczyzn $\Theta^{X\ Y\ (Z)}_{2/1;3/2;(2)} = 2,25$. Jak widać kierunek zależności mierzonej stosunkiem szans jest taki sam, jednak zależność ta jest znacznie silniejsza dla kobiet niż dla mężczyzn.

Pojęcie stosunku szans będzie użyteczne do sformułowania złożonej hipotezy $[XY][XZ][YZ]$. Zgodnie z nią wszystkie pary zmiennych mogą być warunkowo zależne stochastycznie względem trzeciej zmiennej. Jednakże stosunki szans wyznaczone dla każdej pary zmiennych np. $X$ i $Y$ są identyczne dla każdej wartości trzeciej zmiennej. Zmienne są więc od siebie parami zależne, ale zależność ta — mierzona stosunkami szans — jest taka sama w każdej podzbiorowości wyróżnionej przez trzecią zmienną. Formalnie można zapisać to jako:

$$
\begin{aligned}
\Theta^{X\ \ Y\ (Z)}_{a/b;c/d(1)} &= \Theta^{X\ \ Y\ (Z)}_{a/b;c/d(2)} = \ldots = \Theta^{X\ \ Y\ (Z)}_{a/b;c/d(k)} = \ldots = \Theta^{X\ \ Y\ (Z)}_{a/b;c/d(t)}, \\
\Theta^{X\ \ Z\ (Y)}_{e/f;g/h(1)} &= \Theta^{X\ \ Z\ (Y)}_{e/f;g/h(2)} = \ldots = \Theta^{X\ \ Z\ (Y)}_{e/f;g/h(j)} = \ldots = \Theta^{X\ \ Z\ (Y)}_{e/f;g/h(c)}, \\
\Theta^{Y\ \ Z\ (X)}_{m/n;r/s(1)} &= \Theta^{Y\ \ Z\ (X)}_{m/n;r/s(2)} = \ldots = \Theta^{Y\ \ Z\ (X)}_{m/n;r/s(i)} = \ldots = \Theta^{Y\ \ Z\ (X)}_{m/n;r/s(r)},
\end{aligned}
\tag{1.28}
$$

dla dowolnych kategorii poszczególnych zmiennych. Powyższe warunki możemy przełożyć na prawdopodobieństwa rozkładu łącznego. W stosunku do tabeli 1.8 zachodzą następujące warunki:

$$
\frac{\pi^{XYZ}_{111}}{\pi^{XYZ}_{121}} = \frac{\pi^{XYZ}_{112}}{\pi^{XYZ}_{122}} \quad \text{oraz} \quad \frac{\pi^{XYZ}_{111}}{\pi^{XYZ}_{211}} = \frac{\pi^{XYZ}_{121}}{\pi^{XYZ}_{221}} \quad \text{oraz} \quad \frac{\pi^{XYZ}_{111}}{\pi^{XYZ}_{112}} = \frac{\pi^{XYZ}_{211}}{\pi^{XYZ}_{212}}. \tag{1.29}
$$

Oczywiście, relacja ta odnosi się do wszystkich stosunków szans jakie można wyróżnić w tabeli. Tak jak zostało zasygnalizowane, hipoteza ta pod pewnymi względami różni się od hipotez omawianych poprzednio. Do jej sformułowania nie wystarczają pojęcia równomierności i niezależności stochastycznej. Nie można również przedstawić prawdopodobieństwa rozkładu oczekiwanego zgodnego z tą hipotezą za pomocą formuły odwołującej się do prawdopodobieństw brzegowych. Z tego względu nie jest też możliwe zilustrowanie tej hipotezy w sposób analogiczny do tabel 1.9–1.15. Do określenia rozkładu zgodnego z tą hipotezą potrzebne są rozkłady brzegowe wszystkich par zmiennych, tj. $\{XY\}$, $\{XZ\}$ oraz $\{YZ\}$, a także konieczne jest wykorzystanie iteracyjnych metod numerycznych.

To, że warunkowe stosunki szans są sobie równe nie oznacza, że są one równe brzegowemu stosunkowi szans, tj. nie musi zachodzić $\Theta^{X\ \ Y\ (Z)}_{a/b;c/d(k)} = \Theta^{X\ \ Y}_{a/b;c/d;}$. Wynika to z sygnalizowanego już wcześniej twierdzenia o agregacji (collapsibility theorem), które w bardziej ogólnej formie zostało przedstawione w Aneksie. Zgodnie z nim warunkowe stosunki szans między $X$ a $Y$ względem $Z$ są równe brzegowym stosunkom szans jeśli zmienne $X$ i $Z$ są niezależne warunkowo względem $Y$, bądź $Y$ i $Z$ są niezależne warunkowo względem $X$. W omawianej hipotezie $[XY][XZ][YZ]$ żadna z par nie jest warunkowo niezależna stochastycznie względem trzeciej zmiennej. Dlatego, mimo że warunkowe stosunki szans są sobie równe, to mogą się one różnić od zależności wyznaczonej dla rozkładu brzegowego.

### Podsumowanie hipotez dla trzech zmiennych

W tabeli 1.18 mamy podane wszystkie hipotezy jakie zostały sformułowane dla trzech zmiennych. Jak widać, zostały one pogrupowane, gdyż — dla przykładu —trzy hi-

potezy $[X]$, $[Y]$ oraz $[Z]$ są hipotezami tego samego typu, dotyczą jedynie innych zmiennych.

Hipotezy te zostały — o ile to możliwe — uporządkowane: od rozkładu równomierności o najprostszej strukturze, do rozkładu, na strukturę którego rozkładu nie nakładamy żadnych warunków. Uporządkowanie to jest jedynie częściowe, gdyż — jak zostało zauważone wcześniej — nie można orzec, który typ hipotezy $[X][Y][Z]$ czy też $[XY]$ opisuje sytuację prostszą.

Lepszym sposobem na przedstawienie relacji pomiędzy hipotezami jest ich graficzna prezentacja (rysunek 1.1). Hipotezy znajdujące się na tej samej wysokości są hipotezami tego samego typu, dotyczą jedynie innych zmiennych. Strzałki pokazują przejście od hipotezy bardziej skomplikowanej do hipotezy prostszej. Jeśli od jednej hipotezy jedna bądź kilka strzałek prowadzi do innej hipotezy oznacza to, że mogą być one porównywalne ze względu na swoją „prostotę". Jak widać, nie istnieje strzałka łącząca hipotezy $[X][Y][Z]$ oraz $[XY]$.

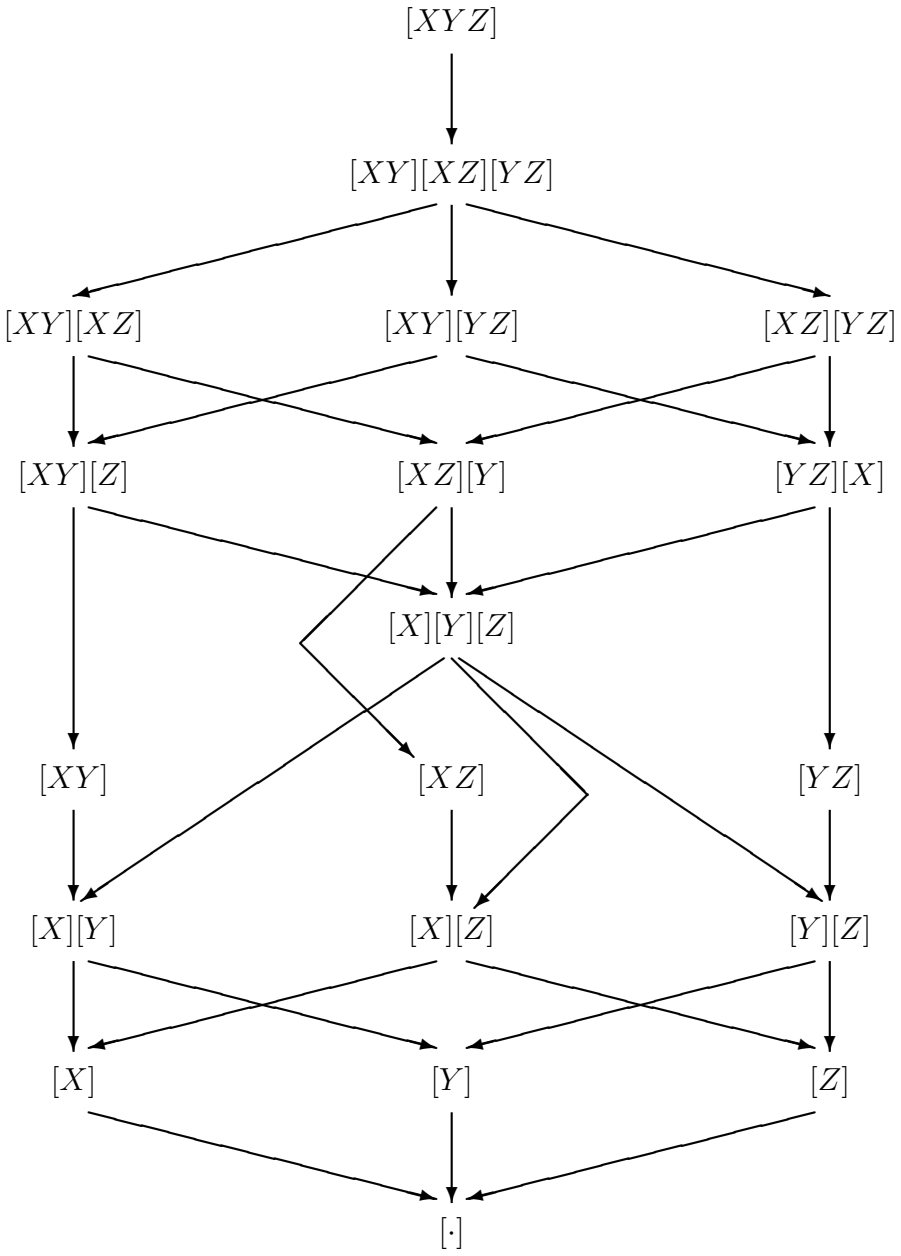
Wszystkie hipotezy — poza $[XY][XZ][YZ]$ — dają opisać się za pomocą pojęć warunkowej lub bezwarunkowej niezależności stochastycznej, bądź równomierności rozkładu. Ponieważ hipoteza $[XY][XZ][YZ]$ opisuje sytuację, w której wszystkie zmienne mogą być parami zależne od siebie, do jej sformułowania konieczne było wprowadzenie pojęcia stosunku szans. Należy jednak zauważyć, że pojęcia szansy i stosunku szans mogą być również wykorzystane do formułowania prostszych hipotez. Na przykład zgodnie z hipotezą równomierności rozkładu łącznego trzech zmiennych wszystkie szanse jakie można wyróżnić dla kategorii zmiennej złożonej $XYZ$ są równe 1. Zgodnie z hipotezą $[YZ]$ warunkowe szanse zmiennej $X$ względem zmiennej złożonej $YZ$ są równe 1. Hipoteza $[XY][Z]$ głosi, że wszystkie warunkowe stosunki szans dla zmiennych $X$ i $Z$ względem $Y$, oraz $Y$ i $Z$ względem $X$ są równe 1. Można również powiedzieć, że warunkowe stosunki szans dla zmiennych $X$ i $Y$ są takie same w każdej podzbiorowości wyróżnionej przez wartości zmiennej $Z$. Twierdzenie o agregacji pozwala dodatkowo stwierdzić, że są one równe odpowiednim stosunkom szans wyznaczonym dla rozkładu brzegowego $\{XY\}$.

Jeżeli w oznaczeniu hipotezy dwie zmienne znajdują się w jednym nawiasie kwadratowym, tj. $[XY]$, opisuje to sytuację, w której zmienne te mogą być zależne stochastycznie. Mówimy wówczas, że pomiędzy zmiennymi zachodzi interakcja *drugiego rzędu* . Oznaczenie $[XYZ]$ oznacza, że związek pomiędzy dwiema zmiennymi może się różnić w pozdbiorowościach wyodrębnionych ze względu na trzecią zmienną. Określa się to jako interakcję *trzeciego rzędu* .

Tabela 1.18: Hipotezy dotyczące rozkładu łącznego trzech zmiennych

| Hipotezy dotyczące rozkładu | Rozkłady brzegowe konieczne do jednoznacznego określenia rozkładu zgodnego z hipotezą |
|---|---|
| 1. Równomierność rozkładu łącznego trzech zmiennych: | |
| $[\cdot]$ | — |
| 2. Warunkowa równomierność rozkładu dwóch zmiennych względem trzeciej zmiennej: | |
| $[X]$ | $\{X\}$ |
| $[Y]$ | $\{Y\}$ |
| $[Z]$ | $\{Z\}$ |
| 3. Warunkowa równomierność rozkładu jednej zmiennej względem kombinacji wartości dwóch pozostałych zmiennych niezależnych stochastycznie: | |
| $[X][Y]$ | $\{X\}$ oraz $\{Y\}$ |
| $[X][Z]$ | $\{X\}$ oraz $\{Z\}$ |
| $[Y][Z]$ | $\{Y\}$ oraz $\{Z\}$ |
| 4. Niezależność stochastyczna trzech zmiennych: | |
| $[X][Y][Z]$ | $\{X\}$, $\{Y\}$ oraz $\{Z\}$ |
| 5. Warunkowa równomierność rozkładu jednej zmiennej względem kombinacji wartości dwóch pozostałych zmiennych: | |
| $[XY]$ | $\{XY\}$ |
| $[XZ]$ | $\{XZ\}$ |
| $[YZ]$ | $\{YZ\}$ |
| 6. Identyczność rozkładu jednej zmiennej względem kombinacji wartości dwóch pozostałych zmiennych: | |
| $[XY][Z]$ | $\{XY\}$ oraz $\{Z\}$ |
| $[XZ][Y]$ | $\{XZ\}$ oraz $\{Y\}$ |
| $[YZ][X]$ | $\{YZ\}$ oraz $\{X\}$ |
| 7. Warunkowa niezależność dwóch zmiennych względem trzeciej zmiennej: | |
| $[XY][XZ]$ | $\{XY\}$ oraz $\{XZ\}$ |
| $[XY][YZ]$ | $\{XY\}$ oraz $\{YZ\}$ |
| $[XZ][YZ]$ | $\{XZ\}$ oraz $\{YZ\}$ |
| 8. Równość warunkowych stosunków szans: | |
| $[XY][XZ][YZ]$ | $\{XY\}$, $\{XZ\}$ oraz $\{YZ\}$ |
| 9. Dowolny rozkład łączny trzech zmiennych — brak hipotezy : | |
| $[XYZ]$ | $\{XYZ\}$ |

Rysunek 1.1: Zależności pomiędzy hipotezami formułowanymi dla trzech zmiennych $X$, $Y$ oraz $Z$



### 1.1.3 Hipotezy dla większej liczby zmiennych

Możliwe jest analizowanie rozkładów łącznych większej liczby zmiennych. Hipotezy te dają się sformułować za pomocą wprowadzonych do tej pory pojęć równomierności, niezależności stochastycznej, stosunków szans. Przypuśćmy, że mamy do czynienia z rozkładem łącznym sześciu zmiennych $A$, $B$, $C$, $D$, $E$, $F$. Dla przykładu, hipotezę $[ABC][DE]$, można sformułować jako połączenie następujących hipotez:

1. Zmienne złożone $ABC$ oraz $DE$ są niezależne stochastycznie.

2. Zmienna $F$ ma rozkład równomierny w każdej podzbiorowości wyróżnionej przez zmienną złożoną $ABCDE$. Wynika z tego również, że obydwie zmienne są niezależne stochastycznie.

W takiej sytuacji dopuszczamy istnienie interakcji drugiego rzędu pomiędzy zmiennymi $D$ oraz $E$, jak również interakcji trzeciego rzędu pomiędzy zmiennymi $A$, $B$ oraz $C$. Jak widać powyższą hipotezę można wypowiedzieć wyłącznie za pomocą za pomocą pojęć warunkowej niezależności i równomierności. Prawdopodobieństwo rozkładu łącznego zgodnego z tą hipotezą jesteśmy w stanie przedstawić jako:

$$\pi^{ABCDEF}_{ijklmn} = \frac{\pi^{ABC}_{ijk} \cdot \pi^{DE}_{lm}}{w} \qquad \text{dla wszystkich kombinacji } i, j, k, l, m, n, \tag{1.30}$$

gdzie $w$ jest liczbą kategorii zmiennej $F$. Do wyznaczenia rozkładu oczekiwanego zgodnego z tą hipotezą wystarczą więc rozkłady brzegowe $\{ABC\}$ oraz $\{DE\}$. Mimo, że hipoteza dotyczy rozkładu aż sześciu zmiennych, do jej wyznaczenia nie są potrzebne procedury iteracyjne.

## 1.2 Parametryzacja modelu logarytmiczno–liniowego

Jak zostało zasygnalizowane na początku tego rozdziału w modelach logarytmiczno–liniowych liczebności lub prawdopodobieństwa rozkładu łącznego przedstawia się w postaci parametrycznej. Parametry opisują rozkład zmiennych i związki pomiędzy nimi. Im prostszą hipotezę opisuje model tym mniej parametrów jest potrzebnych do opisu rozkładu. Poniżej przedstawione zostaną dwie postaci modelu logarytmiczno–liniowego: multiplikatywna i addytywna oraz różne sposoby parametryzacji. Następnie w oparciu o przykładowy rozkład trzech zmiennych, zaprezentowany zostanie sposób interpretowania parametrów modelu.

### 1.2.1 Wersja multiplikatywna i addytywna modelu

Jeśli opisujemy rozkład łączny dwóch zmiennych to prawdopodobieństwo rozkładu łącznego $\pi_{ij}^{XY}$ dla kategorii $i-$tej zmiennej $X$ oraz kategorii $j-$tej zmiennej $Y$ daje się przedstawić jako funkcję następujących parametrów:

$$\pi_{ij}^{XY} = d \cdot d_i^X \cdot d_j^Y \cdot d_{ij}^{XY}. \tag{1.31}$$

Parametr $d$ jest pewną stałą. Parametry $d_i^X$ opisują efekty główne $i-$tej kategorii zmiennej $X$ na prawdopodobieństwo $\pi_{ij}$. W pewnym uproszczeniu można powiedzieć, że wielkość tego parametru zdaje nam sprawę czy kategoria $x_i$ występuje częściej, czy też rzadziej w stosunku innych kategorii zmiennej $X$. Analogicznie parametry $d_j^Y$ opisują efekty główne $j-$tej kategorii zmiennej $Y$. Parametry $d_{ij}^{XY}$ opisują wpływ kombinacji $i-$tej kategorii zmiennej $X$ oraz $j-$tej kategorii zmiennej $Y$, zdają więc sprawę z interakcji pomiędzy zmiennymi. Analogicznie można modelować liczebności rozkładu łącznego.

Powyższy zapis opisuje dowolny rozkład dwóch zmiennych, tj. sytuację, gdy nie mamy żadnej hipotezy odnośnie tego rozkładu. Model taki nazywamy *modelem nasyconym* (*saturated model*). Parametry tego rozkładu pozwalają zrekonstruować dowolny rozkład zmiennych. Zauważmy, że poszczególne parametry korespondują z odpowiednimi rozkładami brzegowymi. Jak zostało powiedziane wcześniej, gdy mamy do czynienia z hipotezą o niezależności stochastycznej do odtworzenia rozkładu wystarczają rozkłady brzegowe obydwu zmiennych tj. $\{X\}$ oraz $\{Y\}$. Dlatego też modelując hipotezę o niezależności stochastycznej wystarczą nam parametry $d$, $d_i^X$ oraz $d_j^Y$ tj.

$$\pi_{ij}^{XY} = d \cdot d_i^X \cdot d_j^Y. \tag{1.32}$$

Innymi słowy zakładamy, że dla każdej kombinacji wartości obydwu zmiennych $X$ oraz $Y$ parametr interakcji $d_{ij}^{XY}$ jest równy 1. W przypadku hipotezy $[X]$ o równomierności rozkładu zmiennej $Y$, zakładamy dodatkowo, że wszystkie parametry $d_j^Y$ są równe 1. Jeżeli mamy do czynienia z hipotezą o równomierności rozkładu łącznego wszystkie parametry $d_i^X$ $d_j^Y$ oraz $d_{ij}^{XY}$ są równe 1. Modele tego typu nazywa się *modelami nienasyconymi*.

Model 1.31 ma postać multiplikatywną tj. parametry są mnożone przez siebie. W literaturze często spotykamy wersję addytywną modelu. Parametry modelu definiują wówczas logarytm prawdopodobieństwa (lub liczebności) rozkładu łącznego. Wówczas

$$\log \pi_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \tag{1.33}$$

gdzie:

$$\mu = \log d, \qquad \lambda_i^X = \log d_i^X, \qquad \lambda_j^Y = \log d_j^Y, \qquad \lambda_{ij}^{XY} = \log d_{ij}^{XY}. \tag{1.34}$$

Od powyższej postaci modele logarytmiczno-liniowe wzięły swoją nazwę. Obie formy: addytywna i multiplikatywna są równoważne. Wersja addytywna pokazuje, że modele logarytmiczno–liniowe są podklasą *uogólnionego modelu liniowego* (Dobson 2002), przy czym zmienną zależną jest logarytm prawdopodobieństwa lub liczebności rozkładu łącznego. Daje się więc pokazać pewne analogie pomiędzy tą metodą a regresją liniową czy analizą wariancji. Uważa się na ogół, że wykonywanie niektórych obliczeń jest wygodniejsze w formie addytywnej. Jednakże interpretacja parametrów modelu w odniesieniu do prawdopodobieństw (liczebności) wydaje się bardziej naturalna niż interpretowanie parametrów w odniesieniu do logarytmów prawdopodobieństw. Dlatego w dalszej części wykorzystywana będzie głównie forma multiplikatywna.

### 1.2.2 Różne wersje parametryzacji

Model 1.31 ma postać ogólną. Istnieje wiele sposobów parametryzacji tego samego modelu. Inaczej mówiąc wielkości $d$, $d_i^X$, $d_j^Y$, $d_{ij}^{XY}$ można zdefiniować na wiele sposobów (Kutylowski 1979, Alba 1987). Istotne jest jednak, aby parametry były zdefiniowane tak, aby posiadały pewną sensowną merytorycznie interpretację i opisywały rozkład zmiennych i związki między zmiennymi.

Wielkości parametrów określane są na podstawie informacji zawartych w rozkładzie. Liczba prawdopodobieństw (liczebności) rozkładu łącznego $\pi_{ij}^{XY}$ jest równa $r \cdot c$, gdzie $r$ oznacza liczbę kategorii zmiennej $X$, a $c$ oznacza liczbę kategorii zmiennej $Y$. Jak widać z równania 1.31 liczba różnych parametrów opisujących rozkład łączny dwóch zmiennych jest większa. Mamy jeden parametr $d$, $r$ parametrów $d_i^X$, $c$ parametrów $d_j^Y$ oraz $c \cdot r$ parametrów $d_{ij}^{XY}$. Oznacza to, że znając wszyskie prawdopodobieństwa $\pi_{ij}$ nie jest możliwe jednoznaczne okreslenie wielkości parametrów bez przyjęcia dodatkowych założeń.

Należy jednak zauważyć, że ważne są nie tyle bezwzględne wartości parametrów, co relacje między nimi. Innymi słowy bardziej aniżeli wielkości parametrów pewnego typu, np. $d_a^X$, $d_b^X$ interesować nas będą relacje pomiędzy nimi, np. wielkość $d_a^X/d_b^X$. Dlatego, sensowne jest przyjęcie warunków, określających relacje między parametrami, które umożliwiają jednoznaczne wyznaczenie ich wielkości. Warunki te można określić na różne sposoby. Można przyjąć, że niektóre wyróżnione parametry są równe pewnej stałej np. 1. Wówczas wartości pozostałych parametrów ustalane są *względem* parametru wyróżnionego. Dla różnych parametryzacji warunki te są inne i co się z tym wiąże odmienna jest interpretacja parametrów. Poniżej przedstawione zostaną dwie parametryzacje najczęściej spotykane w literaturze: parametryzacja *odchyleń*

*multiplikatywnych* (*effect coding*) oraz parametryzacja *względem kategorii odniesienia* (*dummy coding*).

### Parametryzacja odchyleń multiplikatywnych

Aby odróżnić parametryzację odchyleń multiplikatywnych od zapisu ogólnego 1.31, jej parametry oznaczać będziemy jako $\tau$. Tak więc w wersji multiplikatywnej, prawdopodobieństwa modelu nasyconego można przedstawić jako:

$$\pi_{ij}^{XY} = \tau \cdot \tau_i^X \cdot \tau_j^Y \cdot \tau_{ij}^{XY}. \tag{1.35}$$

W parametryzacji odchyleń multiplikatywnych przyjmuje się, że iloczyn parametrów danego typu po jakimkolwiek indeksie dolnym wynosi 1, tj.

$$\prod_{i=1}^{r} \tau_i^X = 1, \qquad \prod_{j=1}^{c} \tau_j^Y = 1, \qquad \prod_{i=1}^{r} \tau_{ij}^{XY} = 1, \qquad \prod_{j=1}^{c} \tau_{ij}^{XY} = 1. \tag{1.36}$$

Po przyjęciu tych założeń, liczba niezależnych parametrów modelu nasyconego jest mniejsza: mamy jeden parametr $\tau$, $(r-1)$ parametrów $\tau_i^X$, $(c-1)$ parametrów $\tau_j^Y$, oraz $(r-1)(c-1)$ parametrów $\tau_{ij}^{XY}$. Liczba niezależnych parametrów wynosi $rc$ i jest równa liczbie prawdopodobieństw rozkładu łącznego. Tak jak zostało powiedziane, hipotezy prostsze można opisywać za pomocą mniejszej liczby parametrów. W przypadku niezależności stochastycznej wystarcza jedynie $(r-1)+(c-1)+1$ niezależnych parametrów.

Przedstawione zostaną teraz formuły dotyczące parametrów modelu nasyconego przy tej parametryzacji. Pomnożenie przez siebie wszystkich $r \cdot c$ prawdopodobieństw modelu 1.35 daje:

$$\prod_{i=1}^{r}\prod_{j=1}^{c} \pi_{ij}^{XY} = \prod_{i=1}^{r}\prod_{j=1}^{c} \tau \cdot \tau_i^X \cdot \tau_j^Y \cdot \tau_{ij}^{XY}. \tag{1.37}$$

Korzystając zaś z warunków ograniczających 1.36 otrzymujemy:

$$\prod_{i=1}^{r}\prod_{j=1}^{c} \pi_{ij}^{XY} = \tau^{rc} \qquad \text{z tego zaś wynika, że} \qquad \tau = \left( \prod_{i=1}^{r}\prod_{j=1}^{c} \pi_{ij}^{XY} \right)^{\frac{1}{rc}}. \tag{1.38}$$

Tak więc parametr $\tau$ jest równy średniej geometrycznej wszystkich prawdopodobieństw rozkładu łącznego. Zapisy 1.37 oraz 1.38 można zobrazować na przykładzie dwóch zmiennych $X$ oraz $Y$ przyjmujących po dwie wartości. Wówczas:

$$\begin{aligned} \prod_{i=1}^{r}\prod_{j=1}^{c} \pi_{ij}^{XY} &= \pi_{11}^{XY}\pi_{12}^{XY}\pi_{21}^{XY}\pi_{22}^{XY} = \\ &= \tau \cdot \tau_1^X \cdot \tau_1^Y \cdot \tau_{11}^{XY} \tau \cdot \tau_1^X \cdot \tau_2^Y \cdot \tau_{12}^{XY} \tau \cdot \tau_2^X \cdot \tau_1^Y \cdot \tau_{21}^{XY} \tau \cdot \tau_2^X \cdot \tau_2^Y \cdot \tau_{22}^{XY.} \end{aligned}$$

a ponieważ z warunków 1.36 wynika, że $\tau_1^X \cdot \tau_2^X = 1$, $\tau_1^Y \cdot \tau_2^Y = 1$, $\tau_{12}^{XY} \cdot \tau_{22}^{XY} = 1$ itd. wobec tego:

$$\prod_{i=1}^{r}\prod_{j=1}^{c} \pi_{ij}^{XY} = \tau^4.$$

Podobnie daje się wyprowadzić wzór na parametr $\tau_i^X$ (oraz $\tau_j^Y$). Mnożąc wszystkie prawdopodobieństwa $\pi_{ij}^{XY}$ po indeksach $j$ (w przypadku parametru $\tau_j^Y$ odpowiednio po indeksach $i$) otrzymujemy:

$$\tau_i^X = \frac{\left(\prod_{j=1}^{c} \pi_{ij}^{XY}\right)^{\frac{1}{c}}}{\tau} \qquad \text{oraz} \qquad \tau_j^Y = \frac{\left(\prod_{i=1}^{r} \pi_{ij}^{XY}\right)^{\frac{1}{r}}}{\tau}. \tag{1.39}$$

Parametr $\tau_i^X$ może być więc interpretowany jako multiplikatywne odchylenie: porównuje się średnią geometryczną prawdopodobieństw $\pi_{ij}^{XY}$ dla $i-$tej wartości zmiennej $X$ do średniej geometrycznej prawdopodobieństw w całym rozkładzie. Można więc interpretować parametry $\tau_i^X$ jako czynniki modyfikujące rozkład łączny obydwu zmiennych ze względu na przynależność do $i-$tej kategorii zmiennej $X$. Znając formuły na parametry $\tau$, $\tau_i^X$ oraz $\tau_j^Y$ bezpośrednio z równania 1.35 można uzyskać formułę na każdy z parametrów $\tau_{ij}^{XY}$:

$$\tau_{ij}^{XY} = \frac{\pi_{ij}^{XY}}{\tau \cdot \tau_i^X \cdot \tau_j^Y} \tag{1.40}$$

Parametr ten zdaje sprawę z multiplikatywnego odchylenia prawdopodobieństwa związanego z przynależnością do kategorii będącej kombinacją wartości dwóch zmiennych tj. $i-$tej kategorii zmiennej $X$ oraz $j-$tej kategorii zmiennej $Y$, od prawdopodobieństwa wynikającego ze średniej ogólnej prawdopodobieństw (czyli $\tau$) zmodyfikowanej przez efekty główne zmiennych $X$ i $Y$ (czyli $\tau_i^X$ jest $\tau_j^Y$). Jeśli parametr ten jest równy 1 oznacza to, że nie istnieje interakcja pomiędzy zmiennymi, tj. zmienne są niezależne. Jeśli istnieje interakcja między zmiennymi oznacza to, że prawdopodobieństwo określone przez parametry $\tau$, $\tau_i^X$ oraz $\tau_j^Y$ powinno być zmodyfikowane przez wielkość $\tau_{ij}^{XY}$. Jeśli jest ona np. większy od 1, wówczas prawdopodobieństwo dla danej kombinacji obydwu zmiennych jest większe niż wynikałoby z parametrów $\tau$, $\tau_i^X$ oraz $\tau_j^Y$, tj. hipotetycznej sytuacji niezależności stochastycznej. Tabela 1.19 stanowi ilustrację modelu nasyconego zgodnego z parametryzacją odchyleń multiplikatywnych dla dwóch zmiennych.

Podobnie parametryzować można model logarytmiczno-liniowy dla większej liczby zmiennych. Jeżeli rozpatrywany jest rozkład łączny trzech zmiennych $X$, $Y$, $Z$ wówczas:

$$\pi_{ijk}^{XYZ} = \tau \cdot \tau_i^X \cdot \tau_j^Y \cdot \tau_k^Z \cdot \tau_{ij}^{XY} \cdot \tau_{ik}^{XZ} \cdot \tau_{jk}^{YZ} \cdot \tau_{ijk}^{XYZ}. \tag{1.41}$$

Tabela 1.19: Rozkład łączny zmiennych $X$ i $Y$ ilustrujący model nasycony zgodny z parametryzacją odchyleń multiplikatywnych

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $x_1$ | $\tau \cdot \tau_1^X \cdot \tau_1^Y \cdot \tau_{11}^{XY}$ | $\tau \cdot \tau_1^X \cdot \tau_2^Y \cdot \tau_{12}^{XY}$ | $\tau \cdot \tau_1^X \cdot \tau_3^Y \cdot \tau_{13}^{XY}$ | $\tau \cdot \tau_1^X \cdot \tau_4^Y \cdot \tau_{14}^{XY}$ |
| $x_2$ | $\tau \cdot \tau_2^X \cdot \tau_1^Y \cdot \tau_{21}^{XY}$ | $\tau \cdot \tau_2^X \cdot \tau_2^Y \cdot \tau_{22}^{XY}$ | $\tau \cdot \tau_2^X \cdot \tau_3^Y \cdot \tau_{23}^{XY}$ | $\tau \cdot \tau_2^X \cdot \tau_4^Y \cdot \tau_{24}^{XY}$ |
| $x_3$ | $\tau \cdot \tau_3^X \cdot \tau_1^Y \cdot \tau_{31}^{XY}$ | $\tau \cdot \tau_3^X \cdot \tau_2^Y \cdot \tau_{32}^{XY}$ | $\tau \cdot \tau_3^X \cdot \tau_3^Y \cdot \tau_{33}^{XY}$ | $\tau \cdot \tau_3^X \cdot \tau_4^Y \cdot \tau_{34}^{XY}$ |
| $x_4$ | $\tau \cdot \tau_4^X \cdot \tau_1^Y \cdot \tau_{41}^{XY}$ | $\tau \cdot \tau_4^X \cdot \tau_2^Y \cdot \tau_{42}^{XY}$ | $\tau \cdot \tau_4^X \cdot \tau_3^Y \cdot \tau_{43}^{XY}$ | $\tau \cdot \tau_4^X \cdot \tau_4^Y \cdot \tau_{44}^{XY}$ |

W równaniu tym pojawił się parametr $\tau_{ijk}^{XYZ}$ opisujący interakcję trzeciego rzędu. Parametry te informują o tym, w jakim stopniu zależność pomiędzy dwiema zmiennymi np. $X$ i $Y$ zależy od wartości trzeciej zmiennej $Z$. Analogicznie do 1.36 zgodnie z parametryzacją względem odchyleń multiplikatywnych przyjmuje się, że:

$$\prod_{i=1}^{r} \tau_i^X = \prod_{j=1}^{c} \tau_j^Y = \prod_{k=1}^{t} \tau_k^Z = 1, \tag{1.42}$$

$$\prod_{i=1}^{r} \tau_{ij}^{XY} = \prod_{j=1}^{c} \tau_{ij}^{XY} = \prod_{i=1}^{r} \tau_{ik}^{XZ} = \prod_{k=1}^{t} \tau_{ik}^{XZ} = \prod_{j=1}^{c} \tau_{jk}^{YZ} = \prod_{k=1}^{t} \tau_{jk}^{YZ} = 1,$$

$$\prod_{i=1}^{r} \tau_{ijk}^{XYZ} = \prod_{j=1}^{c} \tau_{ijk}^{XYZ} = \prod_{k=1}^{t} \tau_{ijk}^{XYZ} = 1.$$

Formuły dotyczące parametrów przy tej parametryzacji będą wynosiły:

$$\tau = \left( \prod_{i=1}^{r} \prod_{j=1}^{c} \prod_{k=1}^{t} \pi_{ijk}^{XYZ} \right)^{\frac{1}{rct}} \tag{1.43}$$

$$\tau_i^X = \frac{\left( \prod_{j=1}^{c} \prod_{k=1}^{t} \pi_{ijk}^{XYZ} \right)^{\frac{1}{ct}}}{\tau},$$

$$\tau_{ij}^{XY} = \frac{\left( \prod_{k=1}^{t} \pi_{ijk}^{XYZ} \right)^{\frac{1}{t}}}{\tau \cdot \tau_i^X \cdot \tau_j^Y},$$

$$\tau_{ijk}^{XYZ} = \frac{\pi_{ijk}}{\tau \cdot \tau_i^X \cdot \tau_j^Y \cdot \tau_k^Z \cdot \tau_{ij}^{XY} \cdot \tau_{ik}^{XZ} \cdot \tau_{jk}^{YZ}}.$$

Parametry $\tau_j^Y, \tau_k^Z, \tau_{ik}^{XZ}, \tau_{jk}^{YZ}$ można określić analogicznie. Jak pokazują powyższe wzory, interpretacja poszczególnych parametrów jest podobna jak w przypadku dwóch

zmiennych. Parametry opisujące interakcję drugiego rzędu $\tau_{ij}^{XY}$ określają na ile średnia geometryczna prawdopodobieństw dla kombinacji $i-$tej wartości zmiennej $X$ oraz $j-$tej wartości zmiennej $Y$ różni się od średniej ogólnej prawdopodobieństw zmodyfikowanej przez efekty główne zmiennych $X$ i $Y$. Natomiast parametry opisujące interakcję trzeciego rzędu $\tau_{ijk}^{XYZ}$ określają w jakim stopniu prawdopodobieństwa związane z przynależnością do $i-$tej wartości zmiennej $X$, $j-$tej wartości zmiennej $Y$ oraz $k-$tej wartości zmiennej $Z$ różnią się od średniej ogólnej prawdopodobieństw ($\tau$) zmodyfikowanej przez efekty główne zmiennych ($\tau_i^X, \tau_j^Y, \tau_k^Z$) oraz interakcje drugiego rzędu ($\tau_{ij}^{XY}, \tau_{ik}^{XZ}, \tau_{jk}^{YZ}$). Jeśli wszystkie parametry tego typu są równe 1 wówczas interakcja trzeciego rzędu nie występuje. Model taki opisuje sytuację, gdy rozkład zgodny jest z hipotezą o równości warunkowych stosunków szans $[XY][XZ][YZ]$.

Daje się również przedstawić związki pomiędzy zdefiniowanymi powyżej parametrami modelu logarytmiczno-liniowego a szansami i stosunkami szans. Można np. pokazać, że iloraz parametrów określający efekty główne zmiennej X dla kategorii $x_a$ oraz $x_b$ tj. $\tau_a^X/\tau_b^X$ jest równy średniej geometrycznej warunkowych szans związanych z tymi kategoriami. Średnią warunkowych szans nazywać będziemy *cząstkową szansą* (*partial odd*) i oznaczać jako $\Omega_{a/b(**)}^{X\ (YZ)}$. Tak więc:

$$\frac{\tau_a^X}{\tau_b^X} = \frac{\left(\prod_{j=1}^{c}\prod_{k=1}^{t}\pi_{ajk}^{XYZ}\right)^{\frac{1}{ct}}}{\left(\prod_{j=1}^{c}\prod_{k=1}^{t}\pi_{bjk}^{XYZ}\right)^{\frac{1}{ct}}} = \left(\prod_{j=1}^{c}\prod_{k=1}^{t}\frac{\pi_{ajk}^{XYZ}}{\pi_{bjk}^{XYZ}}\right)^{\frac{1}{ct}} = \left(\prod_{j=1}^{c}\prod_{k=1}^{t}\Omega_{a/b(jk)}^{X\ (YZ)}\right)^{\frac{1}{ct}} = \Omega_{a/b(**)}^{X\ (YZ)} \tag{1.44}$$

Jak widać mamy do czynienia ze średnią obliczoną ze wszystkich warunkowych szans jakie można wyróżnić ze względu na wartości zmiennej złożonej $YZ$. Tak więc porównanie ze sobą odpowiednich parametrów pierwszego rzędu, daje informację o relacjach pomiędzy warunkowymi prawdopodobieństwami kategorii $x_a$ oraz $x_b$ we wszystkich podzbiorowościach wyróżnionych ze względu pozostałe zmienne.

Można zadać pytanie, czy jedna ze zmiennych np. $Y$ różnicuje określone powyżej cząstkowe szanse $\Omega_{a/b(**)}^{X\ (YZ)}$. Możliwe jest wyznaczenie średniej warunkowych szans dla kategorii $x_a$ oraz $x_b$ w podzbiorowości $Y = y_j$, jakie można wyróżnić w podzbiorowościach zdefiniowanych przez zmienną Z. Zdefiniowana w ten sposób cząstkowa warunkowa szansa będziemy oznaczana $\Omega_{a/b(j*)}^{X\ (YZ)}$. Okazuje się, że istnieje związek pomiędzy tą wielkością a parametrami opisującymi interakcję drugiego rzędu. Mianowicie:

$$\Omega_{a/b(j*)}^{X\ (YZ)} = \left(\prod_{k=1}^{t}\Omega_{a/b(jk)}^{X\ (YZ)}\right)^{\frac{1}{t}} = \left(\prod_{k=1}^{t}\frac{\pi_{ajk}^{XYZ}}{\pi_{bjk}^{XYZ}}\right)^{\frac{1}{t}} = \frac{\tau_a^X \cdot \tau_{aj}^{XY}}{\tau_b^X \cdot \tau_{bj}^{XY}} \tag{1.45}$$

Jeśli natomiast chcemy odtworzyć warunkową szansę w podzbiorowości będącą kombinacją zmiennych $Y$ i $Z$ potrzebne będą parametry opisujące interakcje trzeciego rzędu, tj.

$$\Omega^{X\ (YZ)}_{a/b(jk)} = \frac{\pi^{XYZ}_{ajk}}{\pi^{XYZ}_{bjk}} = \frac{\tau^{X}_{a} \cdot \tau^{XY}_{aj} \cdot \tau^{XZ}_{ak} \cdot \tau^{XYZ}_{ajk}}{\tau^{X}_{b} \cdot \tau^{XY}_{bj} \cdot \tau^{XZ}_{bk} \cdot \tau^{XYZ}_{bjk}} \tag{1.46}$$

Powyższe związki pomiędzy parametrami i warunkowymi szansami pokazują, że odtworzenie bardziej szczegółowych informacji odnośnie rozkładu jednej zmiennej — w tym przypadku zmiennej $X$ — wymaga użycia parametrów opisujących interakcję wyższego rzędu.

Daje się również pokazać relacje pomiędzy parametrami opisującymi interakcję drugiego rzędu a stosunkami szans. Średnią geometryczną warunkowych stosunków szans nazywać będziemy *cząstkowym stosunkiem szans* i oznaczać $\Theta^{X\ \ Y\ (Z)}_{a/b;c/d(*)}$:

$$\Theta^{X\ \ Y\ (Z)}_{a/b;c/d(*)} = \left( \prod_{k=1}^{t} \Theta^{X\ \ Y\ (Z)}_{a/b;c/d(k)} \right)^{\frac{1}{t}} = \left( \prod_{k=1}^{t} \frac{\pi^{XYZ}_{ack} \cdot \pi^{XYZ}_{bdk}}{\pi^{XYZ}_{adk} \cdot \pi^{XYZ}_{bck}} \right)^{\frac{1}{t}} = \frac{\tau^{XY}_{ac} \cdot \tau^{XY}_{bd}}{\tau^{XY}_{ad} \cdot \tau^{XY}_{bc}}. \tag{1.47}$$

Jak widać, parametry opisujące interakcję pomiędzy kategoriami $x_a$, $x_b$ zmiennej $X$ oraz $y_c$, $y_d$ zmiennej $Y$ pozwalają na odtworzenie cząstkowego stosunku szans. Warunkowe stosunki szans, z których obliczana jest średnia wyznaczane są w podzbiorowościach wyróżnionych ze względu na zmienną $Z$. Porównując wzory 1.45 i 1.47, widzimy, że iloraz cząstkowych szans warunkowych daje nam odpowiedni cząstkowy stosunek szans.

$$\Theta^{X\ \ Y\ (Z)}_{a/b;c/d(*)} = \frac{\Omega^{X\ (YZ)}_{a/b(c*)}}{\Omega^{X\ (YZ)}_{a/b(d*)}}. \tag{1.48}$$

Można również pokazać relacje pomiędzy parametrami opisującymi interakcje trzeciego rzędu a warunkowymi stosunkami szans. Mianowicie:

$$\Theta^{X\ \ Y\ (Z)}_{a/b;c/d(k)} = \frac{\Omega^{X\ (YZ)}_{a/b(ck)}}{\Omega^{X\ (YZ)}_{a/b(dk)}} = \frac{\pi^{XYZ}_{ack} \cdot \pi^{XYZ}_{bdk}}{\pi^{XYZ}_{adk} \cdot \pi^{XYZ}_{bck}} = \frac{\tau^{XY}_{ac}\tau^{XY}_{bd}\tau^{XYZ}_{ack}\tau^{XYZ}_{bdk}}{\tau^{XY}_{ad}\tau^{XY}_{bc}\tau^{XYZ}_{adk}\tau^{XYZ}_{bck}} = \Theta^{X\ \ Y\ (Z)}_{a/b;c/d(*)} \cdot \frac{\tau^{XYZ}_{ack}\tau^{XYZ}_{bdk}}{\tau^{XYZ}_{adk}\tau^{XYZ}_{bck}} \tag{1.49}$$

Odtworzenie konkretnego warunkowego stosunku szans jest możliwe przez pomnożenie cząstkowego stosunku szans przez odpowiednie parametry opisujące interakcje trzeciego rzędu.

Możliwe jest oczywiście sformułowanie addytywnej postaci modelu korespondującej z parametryzacją odchyleń multiplikatywnych. Analogiczny do 1.36 warunek ograniczający liczbę niezależnych parametrów, zakłada wówczas, że suma parametrów danego typu po jakimkolwiek indeksie dolnym wynosi 0. Dla modelu opisującego

rozkład dwóch zmiennych:

$$\sum_{i=1}^{r} \lambda_i^X = 0, \qquad \sum_{j=1}^{c} \lambda_j^Y = 0, \qquad \sum_{i=1}^{r} \lambda_{ij}^{XY} = 0, \qquad \sum_{j=1}^{c} \lambda_{ij}^{XY} = 0. \tag{1.50}$$

Formuły dotyczące parametrów przy tej parametryzacji będą wynosiły:

$$\begin{aligned} \mu &= \frac{\left(\sum_{i=1}^{r}\sum_{j=1}^{c} \log \pi_{ij}^{XY}\right)}{rc}, \\ \lambda_i^X &= \frac{\left(\sum_{j=1}^{c} \log \pi_{ij}^{XY}\right)}{c} - \mu, \\ \lambda_j^Y &= \frac{\left(\sum_{i=1}^{r} \log \pi_{ij}^{XY}\right)}{r} - \mu, \\ \lambda_{ij}^{XY} &= \log \pi_{ij}^{XY} - \mu - \lambda_i^X - \lambda_j^Y. \end{aligned} \tag{1.51}$$

Parametr $\mu$ może być więc interpretowany jako średnia arytmetyczna logarytmów prawdopodobieństw. Parametr $\lambda_i^X$ jako różnica między tą średnią a średnią wyliczoną dla danej kategorii zmiennej X. Parametry $\lambda_j^Y$, $\lambda_{ij}^{XY}$ możemy interpretować analogicznie. Podobnie można sformułować addytywną postać modelu dla większej liczby zmiennych.

### Parametryzacja względem kategorii odniesienia

Poniżej przedstawiony zostanie inny sposób parametryzacji spotykany często w literaturze. W parametryzacji *względem kategorii odniesienia* (*dummy effect*) przyjmuje się, że parametry związane z pewnymi ustalonymi wartościami zmiennych są równe 1. Kategorie te nazywać będziemy kategoriami odniesienia. Ich wybór zawsze pozostaje do pewnego stopnia arbitralny. Parametry związane z tą parametryzacją oznaczane będą jako $\gamma$. W wersji multiplikatywnej prawdopodobieństwa modelu nasyconego opisującego rozkład trzech zmiennych $X$,$Y$, $Z$ można przedstawić jako:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ij}^{XY} \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \gamma_{ijk}^{XYZ}. \tag{1.52}$$

Przyjmijmy, ze kategoriami odniesienia dla poszczególnych zmiennych będą: $x_a$, $y_b$ $z_c$. Wówczas zakłada sie, że dowolnych $i$, $j$ oraz $k$:

$$\begin{aligned} \gamma_a^X = \gamma_b^Y = \gamma_c^Z = 1, \\ \gamma_{aj}^{XY} = \gamma_{ak}^{XZ} = \gamma_{ib}^{XY} = \gamma_{bk}^{YZ} = \gamma_{ic}^{XZ} = \gamma_{jc}^{YZ} = 1, \\ \gamma_{ajk}^{XYZ} = \gamma_{ibk}^{XYZ} = \gamma_{ijc}^{XYZ} = 1. \end{aligned} \tag{1.53}$$

Przyjmując powyższe założenia łatwo wyznaczyć parametry modelu dla pozostałych kategorii.

$$\begin{aligned}
\gamma &= \pi^{XYZ}_{abc}, \\
\gamma^X_i &= \frac{\pi^{XYZ}_{ibc}}{\pi^{XYZ}_{abc}} = \Omega^{X\ \ (YZ)}_{i/a;(bc)}, \\
\gamma^{XY}_{ij} &= \frac{\pi^{XYZ}_{ijc}}{\gamma \cdot \gamma^X_i \cdot \gamma^Y_j} = \frac{\pi^{XYZ}_{abc} \cdot \pi^{XYZ}_{ijc}}{\pi^{XYZ}_{ajc} \cdot \pi^{XYZ}_{ibc}} = \Theta^{X\ \ Y\ (Z)}_{i/a;j/b(c)}, \\
\gamma^{XYZ}_{ijk} &= \frac{\pi^{XYZ}_{ijk}}{\gamma \cdot \gamma^X_i \cdot \gamma^Y_j \cdot \gamma^Z_k \cdot \gamma^{XY}_{ij} \cdot \gamma^{XZ}_{ik} \cdot \gamma^{YZ}_{jk}} = \frac{\Theta^{X\ \ Y\ (Z)}_{i/a;j/b(k)}}{\Theta^{X\ \ Y\ (Z)}_{i/a;j/b(c)}}.
\end{aligned} \tag{1.54}$$

Parametry $\gamma^Y_j$, $\gamma^Z_k$, $\gamma^{XZ}_{ik}$, $\gamma^{YZ}_{jk}$ można określić analogicznie. W addytywnej postaci parametryzacji tego typu, zakłada się, że odpowiednie parametry są równe 0.

Powyższe zapisy pokazują relacje pomiędzy parametrami a szansami i stosunkami szans. I tak efekt jednej zmiennej $\gamma^X_i$ posiada interpretacje w postaci szansy kategorii $x_i$ względem kategorii odniesienia $x_a$ wyróżnionej dla kombinacji kategorii odniesienia dwóch pozostałych zmiennych $y_b$ i $z_c$. Podobnie, warunkowe stosunki szans wyróżnione dla kategorii odniesienia można wykorzystać do interpretacji parametrów opisujących interakcję.

Interpretacja ta może wydawać się prostsza w porównaniu do tej związanej parametryzacją odchyleń multiplikatywnych (1.44–1.49). Warto jednak zwrócić uwagę, że o ile parametry $\gamma^X_i$ opisują jedynie rozkład zmiennej $X$ w podzbiorowości będącej kombinacją kategorii odniesienia dwóch pozostałych zmiennych, to parametry $\tau^X_i$ zdają sprawę ze wszystkich rozkładów warunkowych $X$ względem $Y$ i $Z$. Podobnie jest dla parametrów opisujących interakcję wyższego rzędu. Wydaje się, że pod tym względem parametryzacja odchyleń multiplikatywnych pozwala na opis rozkładów zmiennych i związków pomiędzy nimi w bardziej interesujący sposób niż parametryzacja względem parametryzacji względem kategorii odniesienia. Stosowanie tej ostatniej może być jednak wygodne w sytuacji, gdy wyróżnienie jednej kategorii zmiennej jako kategorii odniesienia ma pewne uzasadnienie merytoryczne, ze względu na specyfikę tej kategorii. Tabela 1.20 ilustruje rozkład łączny dwóch zmiennych zgodny z parametryzacją względem kategorii $x_1$, $y_1$. Ponieważ wybrane parametry $\gamma^X_1$, $\gamma^Y_1$, $\gamma^{XY}_{i1}$ $\gamma^{XY}_{1j}$ są równe 1, nie jest konieczne uwzględnienie ich w tej tabeli.

### Interpretacja parametrów modelu nasyconego — przykład

W tej części przedstawiony zostanie przykład interpretacji modelu logarytmiczno-liniowego. Tabela 1.21 przedstawia dane fikcyjne dotyczące trzech zmiennych: uczestnictwa w wyborach parlamentarnych ($X$), wykształcenia ($Y$) oraz miejsca zamiesz-

Tabela 1.20: Rozkład łączny zmiennych $X$ i $Y$ ilustrujący model nasycony zgodny z parametryzacją względem kategorii odniesienia $x_1$, $y_1$

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $x_1$ | $\gamma$ | $\gamma \cdot \gamma_2^Y$ | $\gamma \cdot \gamma_3^Y$ | $\gamma \cdot \gamma_4^Y$ |
| $x_2$ | $\gamma \cdot \gamma_2^X \cdot$ | $\gamma \cdot \gamma_2^X \cdot \gamma_2^Y \cdot \gamma_{22}^{XY}$ | $\gamma \cdot \gamma_2^X \cdot \gamma_3^Y \cdot \gamma_{23}^{XY}$ | $\gamma \cdot \gamma_2^X \cdot \gamma_4^Y \cdot \gamma_{24}^{XY}$ |
| $x_3$ | $\gamma \cdot \gamma_3^X \cdot$ | $\gamma \cdot \gamma_3^X \cdot \gamma_2^Y \cdot \gamma_{32}^{XY}$ | $\gamma \cdot \gamma_3^X \cdot \gamma_3^Y \cdot \gamma_{33}^{XY}$ | $\gamma \cdot \gamma_3^X \cdot \gamma_4^Y \cdot \gamma_{34}^{XY}$ |
| $x_4$ | $\gamma \cdot \gamma_4^X \cdot$ | $\gamma \cdot \gamma_4^X \cdot \gamma_2^Y \cdot \gamma_{42}^{XY}$ | $\gamma \cdot \gamma_4^X \cdot \gamma_3^Y \cdot \gamma_{43}^{XY}$ | $\gamma \cdot \gamma_4^X \cdot \gamma_4^Y \cdot \gamma_{44}^{XY}$ |

kania ($Z$). Rozkład częstości pomiędzy tymi zmiennymi posłużył do wyznaczenia parametrów modelu nasyconego. W tabeli 1.22 zostały zamieszczone parametry zgodne z obydwiema prezentowanymi wcześniej parametryzacjami.

Tabela 1.21: Wykształcenie (Y) a uczestnictwo w wyborach (X) w zależności od miejsca zamieszkania (Z) - rozkład łączny w procentach (dane fikcyjne)

| Wieś ($Z = 1$) | | | |
|---|---|---|---|
| Czy uczestniczył w wyborach? ($X$) | Wykształcenie ($Y$) | | |
| | 1. Podstawowe | 2. Średnie | 3. Wyższe |
| 1. Nie | 6,0 | 4,0 | 2,0 |
| 2. Tak | 5,0 | 5,0 | 3,0 |
| **Małe miasto ($Z = 2$)** | | | |
| Czy uczestniczył w wyborach? ($X$) | Wykształcenie ($Y$) | | |
| | 1. Podstawowe | 2. Średnie | 3. Wyższe |
| 1. Nie | 5,0 | 7,0 | 3,0 |
| 2. Tak | 5,0 | 12,0 | 14,0 |
| **Duże miasto ($Z = 3$)** | | | |
| Czy uczestniczył w wyborach? ($X$) | Wykształcenie ($Y$) | | |
| | 1. Podstawowe | 2. Średnie | 3. Wyższe |
| 1. Nie | 2,0 | 5,0 | 5,0 |
| 2. Tak | 2,0 | 6,0 | 9,0 |

Parametry $\tau$ oraz $\gamma$ zdają jedynie sprawę ze średniej geometrycznej prawdopodobieństw rozkładu łącznego bądź prawdopodobieństwa związanego z kategorią od-

Tabela 1.22: Parametry modelu nasyconego dla danych z tabeli 1.21 (dane fikcyjne)

| Parametryzacja odchyleń multiplikatywnych | | | Parametryzacja względem kategorii odniesienia $x_1$, $y_1$, $z_1$ | | |
|---|---|---|---|---|---|
| efekty główne | interakcje 2–ego rzędu | interakcje 3–ego rzędu | efekty główne | interakcje 2–ego rzędu | interakcje 3–ego rzędu |
| $\tau = 0,048$ | $\tau_{11}^{XY} = 1,238$ | $\tau_{111}^{XYZ} = 0,953$ | $\gamma = 0,060$ | $\gamma_{11}^{XY} = 1,000$ | $\gamma_{111}^{XYZ} = 1,000$ |
| | $\tau_{12}^{XY} = 1,026$ | $\tau_{112}^{XYZ} = 1,142$ | | $\gamma_{12}^{XY} = 1,000$ | $\gamma_{112}^{XYZ} = 1,000$ |
| $\tau_1^X = 0,833$ | $\tau_{13}^{XY} = 0,787$ | $\tau_{113}^{XYZ} = 0,918$ | $\gamma_1^X = 1,000$ | $\gamma_{13}^{XY} = 1,000$ | $\gamma_{113}^{XYZ} = 1,000$ |
| $\tau_2^X = 1,201$ | $\tau_{21}^{XY} = 0,808$ | $\tau_{121}^{XYZ} = 0,939$ | $\gamma_2^X = 0,833$ | $\gamma_{21}^{XY} = 1,000$ | $\gamma_{121}^{XYZ} = 1,000$ |
| | $\tau_{22}^{XY} = 0,975$ | $\tau_{122}^{XYZ} = 1,053$ | | $\gamma_{22}^{XY} = 1,500$ | $\gamma_{122}^{XYZ} = 1,000$ |
| $\tau_1^Y = 0,794$ | $\tau_{23}^{XY} = 1,270$ | $\tau_{123}^{XYZ} = 1,012$ | $\gamma_1^Y = 1,000$ | $\gamma_{23}^{XY} = 1,800$ | $\gamma_{123}^{XYZ} = 1,000$ |
| $\tau_2^Y = 1,271$ | | $\tau_{131}^{XYZ} = 1,117$ | $\gamma_2^Y = 0,666$ | | $\gamma_{131}^{XYZ} = 1,000$ |
| $\tau_3^Y = 0,991$ | $\tau_{11}^{XZ} = 1,115$ | $\tau_{132}^{XYZ} = 0,832$ | $\gamma_3^Y = 0,333$ | $\gamma_{11}^{XZ} = 1,000$ | $\gamma_{132}^{XYZ} = 1,000$ |
| | $\tau_{12}^{XZ} = 0,849$ | $\tau_{133}^{XYZ} = 1,076$ | | $\gamma_{12}^{XZ} = 1,000$ | $\gamma_{133}^{XYZ} = 1,000$ |
| $\tau_1^Z = 0,819$ | $\tau_{13}^{XZ} = 1,056$ | $\tau_{211}^{XYZ} = 1,049$ | $\gamma_1^Z = 1,000$ | $\gamma_{13}^{XZ} = 1,000$ | $\gamma_{211}^{XYZ} = 1,000$ |
| $\tau_2^Z = 1,395$ | $\tau_{21}^{XZ} = 0,897$ | $\tau_{212}^{XYZ} = 0,875$ | $\gamma_2^Z = 0,833$ | $\gamma_{21}^{XZ} = 1,000$ | $\gamma_{212}^{XYZ} = 1,000$ |
| $\tau_3^Z = 0,876$ | $\tau_{22}^{XZ} = 1,178$ | $\tau_{213}^{XYZ} = 1,089$ | $\gamma_3^Z = 0,333$ | $\gamma_{22}^{XZ} = 1,200$ | $\gamma_{213}^{XYZ} = 1,000$ |
| | $\tau_{23}^{XZ} = 0,947$ | $\tau_{221}^{XYZ} = 1,065$ | | $\gamma_{23}^{XZ} = 1,200$ | $\gamma_{221}^{XYZ} = 1,000$ |
| | | $\tau_{222}^{XYZ} = 0,950$ | | | $\gamma_{222}^{XYZ} = 1,142$ |
| | $\tau_{11}^{YZ} = 1,762$ | $\tau_{223}^{XYZ} = 0,989$ | | $\gamma_{11}^{YZ} = 1,000$ | $\gamma_{223}^{XYZ} = 0,800$ |
| | $\tau_{12}^{YZ} = 0,944$ | $\tau_{231}^{XYZ} = 0,895$ | | $\gamma_{12}^{YZ} = 1,000$ | $\gamma_{231}^{XYZ} = 1,000$ |
| | $\tau_{13}^{YZ} = 0,601$ | $\tau_{232}^{XYZ} = 1,203$ | | $\gamma_{13}^{YZ} = 1,000$ | $\gamma_{232}^{XYZ} = 2,592$ |
| | $\tau_{21}^{YZ} = 0,899$ | $\tau_{233}^{XYZ} = 0,929$ | | $\gamma_{21}^{YZ} = 1,000$ | $\gamma_{233}^{XYZ} = 1,000$ |
| | $\tau_{22}^{YZ} = 1,081$ | | | $\gamma_{22}^{YZ} = 2,100$ | |
| | $\tau_{23}^{YZ} = 1,029$ | | | $\gamma_{23}^{YZ} = 3,750$ | |
| | $\tau_{31}^{YZ} = 0,631$ | | | $\gamma_{31}^{YZ} = 1,000$ | |
| | $\tau_{32}^{YZ} = 0,980$ | | | $\gamma_{32}^{YZ} = 1,800$ | |
| | $\tau_{33}^{YZ} = 1,616$ | | | $\gamma_{33}^{YZ} = 7,500$ | |

niesienia, nie są one na ogół celem zainteresowania badacza[4]. Parametry $\tau_i^X$, $\tau_j^Y$, $\tau_k^Z$ opisują rozkłady poszczególnych zmiennych. Dla przykładu: zgodnie ze wzorem 1.44, cząstkowa szansa wyróżniona dla dwóch kategorii zmiennej $Z$ *małe miasto* i *wieś* wynosi

$$\Omega_{2/1(**)}^{Z\ (YX)} = \tau_2^Z/\tau_1^Z = 1,39/0,82 = 1,70.$$

W pewnym uproszczeniu można powiedzieć, że na 7 osób mieszkających na wsi przypada około 12 osób mieszkających w małym mieście. Należy jednak pamiętać, że jest to średnia geometryczna warunkowych szans obliczonych dla wszystkich (sześciu) podzbiorowości wyróżnionych ze względu na wykształcenie i udział w wyborach. Dla rozkładu brzegowego proporcja ta nie musi wynosić tyle samo, w omawianym przypadku wynosi 1,88. Podobnie daje się wyróżnić cząstkowe szanse porównujące proporcje osób mieszkających w dużym mieście do osób mieszkających na wsi oraz proporcję osób mieszkających w dużym mieście do osób mieszkających w małym mieście. Wynoszą one odpowiednio: $\tau_3^Z/\tau_1^Z = 1,07$ oraz $\tau_3^Z/\tau_2^Z = 0,62$.

Podobnie interpretować można szanse związane ze zmienną opisującą wykształcenie badanych. Wyrażenie $\tau_2^Y/\tau_1^Y$=1,60, czyli cząstkowa szansa $\Omega_{2/1(**)}^{Y\ (XZ)}$ pozwala porównać prawdopodobieństwo spotkania osoby ze średnim wykształceniem względem prawdopodobieństwa spotkania osoby z wykształceniem podstawowym. Jest to średnia warunkowych szans wyznaczona dla wszystkich podzbiorowości jakie można wyróżnić ze względu na miejsce zamieszkania i udział w wyborach. Można zadać kolejne pytanie: na ile miejsce zamieszkania różnicuje tę szanse. Aby na nie odpowiedzieć należy wyznaczyć cząstkowe szanse $\Omega_{2/1(*1)}^{Y\ (XZ)}$, $\Omega_{2/1(*2)}^{Y\ (XZ)}$, $\Omega_{2/1(*3)}^{Y\ (XZ)}$. Jak pokazuje równanie 1.45 cząstkowe warunkowe szanse wynoszą odpowiednio:

$$1,60 \cdot \frac{\tau_{21}^{YZ}}{\tau_{11}^{YZ}} = 0,82$$

dla wsi, 1,83 dla małych miast i 2,74 dla dużych miast. Warto zwrócić uwagę, że w tym przypadku są to średnie warunkowych szans wyznaczonych w podzbiorowościach określonych przez dwie kategorie zmiennej *uczestnictwo w wyborach*. Analogicznie cząstkowe szanse porównujące wykształcenie wyższe i średnie wynoszą $\Omega_{3/2(*1)}^{Y\ (XZ)} = 0,54$, dla wsi, $\Omega_{3/2(*2)}^{Y\ (XZ)} = 0,70$ dla małych miast oraz $\Omega_{3/2(*3)}^{Y\ (XZ)} = 1,22$ dla dużych miast. Średnia geometryczna tych szans wynosi $\Omega_{3/2(**)}^{Y\ (XZ)} = \tau_3^Y/\tau_2^Y = 0,77$. Wartości analizowanych powyżej parametrów pokazują, że zmienne miejsce zamieszkania i wykształcenie są w opisywanej zbiorowości zależne warunkowo względem trzeciej zmiennej.

[4]W przypadku, jeśli parametryzujemy rozkład łączny liczebności — a nie prawdopodobieństwa — parametry te zdają sprawę również z liczebności próby lub liczebności kategorii odniesienia. Wartości pozostałych parametrów są oczywiście takie same bez względu na to czy analizujemy rozkład prawdopodobieństwa czy liczebności.

Cząstkowa szansa związana z uczestnictwem wyborczym wynosi $\Omega^{X\ (YZ)}_{2/1(**)} = \tau^X_2/\tau^X_1 = (\tau^X_2)^2 = 1,44$. Wskazuje to, że na dwie osoby, które nie głosowały, przypadają prawie trzy osoby, które wzięły udział w wyborach, przy czym jest to średnia wyznaczona względem pozostałych zmiennych. Można oczekiwać, że wykształcenie będzie te wielkości różnicować. Średnia warunkowa szansa dla osób z wykształceniem podstawowym wynosi:

$$\Omega^{X\ (YZ)}_{2/1(1*)} = 1,44 \cdot \frac{\tau^{XY}_{21}}{\tau^{XY}_{11}} = 1,44 \cdot (\tau^{XY}_{21})^2 = 0,94.$$

Dla wykształcenia średniego i wyższego cząstkowe szanse warunkowe wynoszą odpowiednio 1,37 oraz 2,32. Jak widać, zmienne te są od siebie zależne. Zgodnie z równaniem 1.48 można wyznaczyć cząstkowe stosunki szans. Jeśli porównujemy uczestnictwo wyborcze w kategoriach wykształcenia średniego i podstawowego to wielkość ta wynosi:

$$\Theta^{X\ \ Y\ (Z)}_{2/1;2/1(*)} = \frac{\Omega^{X\ (YZ)}_{2/1(2*)}}{\Omega^{X\ (YZ)}_{2/1(1*)}} = 1,46.$$

Jeśli porównujemy skrajne kategorie wykształcenia otrzymujemy$\Theta^{X\ \ Y\ (Z)}_{2/1;3/1(*)} = 2,47$. Wielkości te opisują średnie warunkowych zależności pomiędzy wykształceniem i uczestnictwem wyborczym wyznaczonych w podzbiorowościach wyróżnionych ze względu na miejsce zamieszkania.

Można zadać pytanie, czy miejsce zamieszkania różnicuje tę zależność, np. czy warunkowe stosunki szans $\Theta^{X\ \ Y\ (Z)}_{2/1;2/1(1)}$, $\Theta^{X\ \ Y\ (Z)}_{2/1;2/1(2)}$, $\Theta^{X\ \ Y\ (Z)}_{2/1;2/1(3)}$, są różne czy też takie same? Innymi słowy pytanie dotyczy tego, czy mamy do czynienia z interakcją trzeciego rzędu. Przytoczone wcześniej twierdzenie o agregacji wskazuje pośrednio, że gdyby wykształcenie i miejsce zamieszkania były niezależne warunkowo, wówczas określone powyżej stosunki szans między wykształceniem a uczestnictwem wyborczym byłyby takie same bez względu na miejsce zamieszkania. Ponieważ jednak — jak zostało pokazane wcześniej - zmienne te są zależne, należy oczekiwać, że miejsce zamieszkania będzie różnicowało zależności pomiędzy wykształceniem i uczestnictwem w głosowaniu. Zgodnie z równaniem 1.49 warunkowe stosunki szans porównujące wykształcenie podstawowe i średnie oraz uczestnictwo wyborcze są równe odpowiednio:

$$\Theta^{X\ \ Y\ (Z)}_{2/1;2/1(1)} = 1,46 \cdot \frac{\tau^{XYZ}_{111} \cdot \tau^{XYZ}_{221}}{\tau^{XYZ}_{121} \cdot \tau^{XYZ}_{211}} = 1,50$$

dla wsi, $\Theta^{X\ \ Y\ (Z)}_{2/1;2/1(2)} = 1,71$ dla małego miasta i $\Theta^{X\ \ Y\ (Z)}_{2/1;2/1(3)} = 1,20$ dla dużego miasta. Analogicznie można porównać stosunki szans uwzględniające skrajne kategorie wykształcenia. Otrzymujemy wówczas wielkości $\Theta^{X\ \ Y\ (Z)}_{2/1;3/1(1)} = 1,80$ dla wsi, $\Theta^{X\ \ Y\ (Z)}_{2/1;3/1(2)} = 4,66$ dla małego miasta i $\Theta^{X\ \ Y\ (Z)}_{2/1;3/1(3)} = 1,80$ dla dużego miasta. W podanym przykładzie, związki pomiędzy wykształceniem a uczestnictwem wyborczym

wydają się znaczące dla każdej kategorii miejsca zamieszkania i kierunek tej zależności jest pozytywny tj. osoby z wyższym wykształceniem głosują częściej (wszystkie stosunki szans podane powyżej są większe od 1). Jednak należy zauważyć, że miejsce zamieszkania różnicuje siłę tego związku. W małym mieście związek ten jest zdecydowanie najsilniejszy. Na wsi i w dużym mieście siła związku jest taka sama w odniesieniu do kategorii skrajnych wykształcenia, jednak można zaobserwować różnice przy porównywaniu stosunków szans uwzględniających wykształcenie podstawowe i średnie Można więc mówić o interakcji trzeciego rzędu pomiędzy opisanymi zmiennymi. Podobnie rozpatrywać można, na ile wykształcenie różnicuje siłę związku pomiędzy uczestnictwem wyborczym a miejscem zamieszkania.

Tak jak zostało zasygnalizowane wcześniej, interpretacja parametrów związanych z drugim rodzajem parametryzacji jest prostsza, ale pod pewnymi względami bardziej ograniczona. Dla przykładu: efekt główny związany z uczestnictwem wyborczym $\gamma_2^X = 0,833$ informuje nas jedynie, że wśród osób z wykształceniem podstawowym mieszkających na wsi na cztery osoby głosujące przypada około pięciu nie biorących udziału wyborach. Jak pokazuje równanie 1.46 informację tę stosunkowo łatwo odtworzyć za pomocą parametrów opisujących multiplikatywne odchylenia:

$$\gamma_2^X = \frac{\tau_2^X \cdot \tau_{21}^{XY} \cdot \tau_{21}^{XZ} \cdot \tau_{211}^{XYZ}}{\tau_1^X \cdot \tau_{11}^{XY} \cdot \tau_{11}^{XZ} \cdot \tau_{111}^{XYZ}}.$$

Wielkość parametru $\gamma_{23}^{XY} = 1,80$ określająca warunkowy stosunek szans dla mieszkańców wsi $\Theta_{2/1;3/1(1)}^{X \quad Y \ (Z)}$, również została przedstawiona powyżej jako funkcja odpowiednich parametrów $\tau$. Parametry $\gamma$ mają bezpośrednią interpretację w postaci warunkowych szans i stosunków szans wyróżnionych dla kategorii odniesienia, jednak odtworzenie na ich podstawie informacji o szansach i stosunkach szans dla pozostałych kategorii jest trudniejsze.

### 1.2.3 Modele hierarchiczne i niehierarchiczne

Jak zostało powiedziane model nasycony opisuje sytuację, w której nie przyjmuje się żadnych założeń dotyczących rozkładu łącznego zmiennych. Modelowanie logarytmiczno–liniowe pozwala nakładać na poszczególne parametry kolejne ograniczenia, aby powstały w ten sposób model nienasycony opisywał hipotezę prostszą. Jednym ze sposobów upraszania modelu jest założenie, że wybrane parametry modelu są równe 1 (lub 0 w wersji addytywnej). Jeżeli rozpatrujemy rozkład łączny pięciu zmiennych $A$, $B$, $C$, $D$, $E$ to hipotezę $[ABC]$ $[AD]$ można modelować jako:

$$\pi_{ijklm}^{ABCDE} = d \cdot d_i^A \cdot d_j^B \cdot d_k^C \cdot d_l^D \cdot d_{ij}^{AB} \cdot d_{ik}^{AC} \cdot d_{jk}^{BC} \cdot d_{il}^{AD} \cdot d_{ijk}^{ABC} \tag{1.55}$$

Jak widać parametry $d_{jl}^{BD}$, $d_{kl}^{CD}$, $d_{ijl}^{ABD}$, $d_{ikl}^{ACD}$, $d_{jkl}^{BCD}$ oraz wszystkie parametry związane ze zmienną $E$ są równe 1.

Innym rodzajem ograniczeń nakładanych na model, może być założenie o równości wybranych parametrów, np. $d_1^X = d_2^X$ co oznacza, że efekty główne związane z pierwszą i druga kategorią zmiennej $X$ są sobie równe.

Przedstawiane do tej pory modele posiadają pewną specyficzną strukturę. Zauważmy, że jeśli w modelu występował parametr opisujący interakcję wyższego rzędu dla określonych zmiennych pociągało to za sobą występowanie parametrów niższego rzędu związanych z tymi zmiennymi. Np. w powyższym modelu 1.55 występowanie parametru pisującego interakcję trzeciego rzędu $d_{ijk}^{ABC}$ pociąga za sobą występowanie parametrów drugiego rzędu $d_{ij}^{AB}$, $d_{ik}^{AC}$ , $d_{jk}^{BC}$ oraz parametrów opisujących efekty główne, tj. $d_i^A$, $d_j^B$, $d_k^C$. Modele tego typu nazywa się *modelami hierarchicznymi*. Ta cecha modeli hierarchicznych umożliwia uproszczony zapis tych modeli. Z zapisu $[ABC]$ $[AD]$ wnioskujemy, że w modelu występują nie tylko parametry $d_{ijk}^{ABC}$ oraz $d_{il}^{AD}$ ale też parametry $d_i^A$, $d_j^B$, $d_k^C$, $d_l^D$ $d_{ij}^{AB}$ $d_{ik}^{AC}$ $d_{jk}^{BC}$. Możliwe jest jednak formułowanie modeli nie–hierarchicznych. Przykładem takim jest model:

$$\pi_{ijklm}^{ABCDE} = d \cdot d_j^B \cdot d_k^C \cdot d_l^D \cdot d_{ij}^{AB} \cdot d_{ik}^{AC} \cdot d_{il}^{AD} \cdot d_{ijk}^{ABC} \tag{1.56}$$

Zauważmy, że w modelu nie występuje efekt główny $d_i^A$ mimo, że występują parametry wyższego rzędu związane ze zmienną $A$. Tak więc zgodnie z tym modelem rozkład zmiennej $A$ jest równomierny, jednak nie zakładamy, że zmienne $A$ i $B$ są niezależne warunkowo względem pozostałych zmiennych, nie czynimy również analogicznego założenia w odniesieniu do par $A$ i $C$ oraz $A$ i $D$. Podobnie, nie występuje również parametr drugiego rzędu $d_{jk}^{BC}$ mimo, że występuje parametr trzeciego rzędu $d_{ijk}^{ABC}$. Modele nie–hierarchiczne są stosunkowo rzadko wykorzystywane w praktyce. Wynika to między innymi z tego, że nie są one łatwe do interpretacji i estymacji. W tej pracy skupiać się będę przede wszystkim na modelach hierarchicznych.

## 1.3 Estymacja i weryfikacja modeli

Do tej pory zaprezentowane zostały poszczególne hipotezy, jak również sposoby ich parametryzacji. W tej części omówione zostaną kwestie estymacji i weryfikacji modeli logarytmiczno–liniowych. Estymacja dotyczy oszacowania na podstawie danych z próby rozkładu zgodnego z testowaną hipotezą. Weryfikacja hipotez polega na konfrontacji tego rozkładu z danymi empirycznymi, a dokładniej na porównaniu liczebności rozkładu zgodnego z daną hipotezą — czyli tzw. rozkładu oczekiwanego — z rozkładem liczebności uzyskanym w próbie losowej, nazywanym rozkładem empirycznym

lub obserwowanym. Liczebności oczekiwane oznaczać będziemy $F_{ij}^{XY} = n \cdot \pi_{ij}^{XY}$, gdzie $n$ oznacza ogólną liczebność próby a $\pi_{ij}^{XY}$ prawdopodobieństwa zgodne z daną hipotezą. Na ogół nie dysponujemy rozkładem oczekiwanym, jak sygnalizowaliśmy podczas prezentacji hipotez nie wynika on bezpośrednio z treści hipotezy, a do jego wyznaczenia potrzebne są dodatkowe informacje o rozkładzie poszczególnych zmiennych, które szacuje się na podstawie próby. Takie oszacowane liczebności i prawdopodobieństwa oczekiwane oznaczymy jako $\widehat{F}_{ij}^{XY}$ oraz $\widehat{\pi}_{ij}^{XY}$. Odpowiednio przez $f_{ij}^{XY}$ oznaczamy liczebności obserwowane w próbie, a przez $p_{ij}^{XY}$ – częstości w próbie, tj. $p_{ij}^{XY} = f_{ij}^{XY}/n$. W przypadku większej liczby zmiennych zastosowane będą analogiczne oznaczenia.

### 1.3.1 Estymacja

Jak pokazuje tabela 1.18, spośród hipotez omawianych do tej pory, jedynie rozkład oczekiwany zgodny z hipotezą $[\cdot]$ można wyznaczyć bezpośrednio: trzeba jedynie znać liczbę kategorii poszczególnych zmiennych i liczebność próby. Dla pozostałych hipotez potrzebna jest informacja dotycząca rozkładów poszczególnych zmiennych, bądź rozkładu łącznego kilku zmiennych. Informacje te estymujemy na podstawie próby, tj. wartość parametru populacyjnego[5] $\vartheta$ — w tym przypadku odpowiedniego prawdopodobieństwa — szacujemy na podstawie statystyki z $n$–elementowej próby, tj. estymatora $T_n$. W modelach logarytmiczno–liniowych wykorzystuje się estymację metodą *największej wiarygodności* (*maximum likelihood estimation*). Estymatorem największej wiarygodności parametru jest wielkość $T_n^{ML}$, o ile spełniony jest warunek:

$$P(\text{obserwowane dane}|\vartheta = T_n^{ML}) \geq P(\text{obserwowane dane}|\vartheta = T_n) \tag{1.57}$$

dla każdej wartości $T_n$. Innymi słowy szukamy takiej wielkości $T_n$, która maksymalizuje prawdopodobieństwo otrzymania obserwowanych danych. Przypuśćmy, że rozważamy hipotezy $[X]$ oraz $[\cdot]$ dla rozkładu dwóch dwuwartościowych zmiennych $X$ oraz $Y$. Hipotezy te chcemy zweryfikować na podstawie danych z tabeli 1.23a. Dla prostoty obliczeń załóżmy, że próba była dobrana w sposób prosty niezależny. Rozkłady prawdopodobieństwa związane z hipotezami $[X]$ oraz $[\cdot]$ przedstawiają tabele 1.23b oraz 1.23c. Rozkład dla hipotezy głoszącej równomierność rozkładu łącznego jest określony jednoznacznie, natomiast dla hipotezy $[X]$ wielkość $\pi_1^X$ jest nieznana. Jeśli chcemy oszacować tę wielkość metodą największej wiarygodności, powinniśmy odpowiedzieć na pytanie: dla jakiej wielkości $\pi_1^X$ najbardziej jest prawdopodobne uzyskanie takiego wyniku jaki otrzymano w próbie.

[5]Słowa parametr używamy tu w znaczeniu bardziej ogólnym niż w części dotyczącej parametryzacji modelu logarytmiczno–liniowego.

Tabela 1.23: Rozkład empiryczny liczebności i rozkłady prawdopodobieństwa zgodne z hipotezami $[X]$ oraz $[\cdot]$

Tabela 1.23a

| X\Y | $y_1$ | $y_2$ | $\Sigma$ |
|---|---|---|---|
| $x_1$ | 1 | 2 | 3 |
| $x_2$ | 4 | 5 | 9 |
| $\Sigma$ | 5 | 7 | 12 |

Tabela 1.23b

| X\Y | $y_1$ | $y_2$ | $\Sigma$ |
|---|---|---|---|
| $x_1$ | $0{,}5\pi_1^X$ | $0{,}5\pi_1^X$ | $\pi_1^X$ |
| $x_2$ | $0{,}5(1\text{-}\pi_1^X)$ | $0{,}5(1\text{-}\pi_1^X)$ | $1-\pi_1^X$ |
| $\Sigma$ | 0,5 | 0,5 | 1 |

Tabela 1.23c

| X\Y | $y_1$ | $y_2$ | $\Sigma$ |
|---|---|---|---|
| $x_1$ | 0,25 | 0,25 | 0,5 |
| $x_1$ | 0,25 | 0,25 | 0,5 |
| $\Sigma$ | 0,5 | 0,5 | 1 |

Prawdopodobieństwo uzyskania danego wyniku z próby, tj. liczebności $f_{ij}^{XY}$ przy założeniu o prawdziwości hipotezy $H_A$, określa tzw. funkcja wiarygodności. Przez $\ell_A$ będziemy oznaczać wartość tej funkcji dla hipotezy A. Prawdopodobieństwo to zależy od tego, z jakim schematem doboru próby mamy do czynienia. Poniżej zakładamy, że mamy do czynienia z doborem prostym niezależnym. Bardziej skomplikowane schematy — stosowane powszechnie — wymagałyby pewnych modyfikacji w stosunku do opisanej poniżej procedury. W tej pracy ograniczę się jedynie do pokazanie głównej idei estymacji i weryfikacji hipotez. Przy przyjętym założeniu, prawdopodobieństwo $\ell_A$, określa formuła[6]:

$$\ell_A = P(f_{11}^{XY}, f_{12}^{XY}, \dots f_{ij}^{XY}, \dots f_{rc}^{XY}|H_A) = \tag{1.58}$$
$$= \frac{n!}{f_{11}^{XY}!f_{12}^{XY}!\dots f_{ij}^{XY}!\dots f_{rc}^{XY}!}(\pi_{11A}^{XY})^{f_{11}^{XY}}(\pi_{12A}^{XY})^{f_{12}^{XY}}\dots(\pi_{ijA}^{XY})^{f_{ij}^{XY}}\dots(\pi_{rcA}^{XY})^{f_{rc}^{XY}}.$$

gdzie $\pi_{ijA}^{XY}$ są prawdopodobieństwami oczekiwanymi zgodnymi z hipotezą $H_A$. Dla danych z tabeli 1.23a, i hipotezy $[X]$ prawdopodobieństwo to wynosi:

$$\ell_{[X]} = \frac{12!}{2!\cdot 4!\cdot 5!}\cdot 0,5^{12}\cdot(\pi_1^X)^3\cdot(1-\pi_1^X)^9 \tag{1.59}$$

Jak łatwo zauważyć, aby powyższe prawdopodobieństwo było możliwie największe, należy zmaksymalizować wielkość $(\pi_1^X)^3\cdot(1-\pi_1^X)^9$, można bowiem zauważyć, że wyrażenie $\frac{n!}{f_{11}^{XY}!f_{12}^{XY}!\dots f_{ij}^{XY}!\dots f_{rc}^{XY}!}$ nie jest związane z parametrem, który chcemy oszacować $(\pi_1^X)$. Maksymalizacja interesującego nas wyrażenia jest równoznaczna z maksy-

[6]Formuła ta jest związana z rozkładem wielomianowym, czyli sytuacją doboru prostego niezależnego. Gdyby rozkład obejmował jedynie dwie komórki, mielibyśmy do czynienia z rozkładem dwumianowym, czyli odpowiednie prawdopodobieństwo można byłoby obliczyć za pomocą schematu Bernoulliego (widoczne są analogie pomiędzy tymi wzorami.)

malizacją jego logarytmu, tj.

$$\log[(\pi_1^X)^3 \cdot (1-\pi_1^X)^9] = \log(\pi_1^X)^3 + \log(1-\pi_1^X)^9 = 3\log(\pi_1^X) + 9\log(1-\pi_1^X)$$

Obliczając pochodną powyższego wyrażenia względem $\pi_1^X$ i przyrównując ją do 0, otrzymujemy:

$$\frac{3}{\pi_1^X} - \frac{9}{1-\pi_1^X} = 0.$$

Rozwiązując powyższe równanie okazuje się, że rozwiązaniem będzie wartość 0,25. Uogólniając oszacowaniem największej wiarygodności będzie odpowiednia częstość brzegowa w próbie, tj. $\widehat{\pi}_1^X = p_1^X = f_i^X/n$.

Podobnie jest dla innych hipotez. Na przykład, jeśli przedmiotem analizy jest rozkład trzech zmiennych i testujemy hipotezę $[X][YZ]$, formuła 1.22 pokazuje, że w celu wyznaczenia rozkładu oczekiwanego konieczne jest oszacowanie wielkości $\pi_i^X$ oraz $\pi_{jk}^{YZ}$. Tak jak poprzednio, oszacowaniem największej wiarygodności są odpowiednie częstości brzegowe w próbie tj.

$$\widehat{\pi}_i^X = p_i^X \qquad \text{oraz} \qquad \widehat{\pi}_{jk}^{YZ} = p_{jk}^{YZ}$$

Jak zostało zasygnalizowane wcześniej, dla pewnych hipotez nie istnieje formuła, pozwalająca wyznaczyć prawdopodobieństwa rozkładu łącznego. Spośród modeli omawianych do tej pory, przykładem takiej hipotezy jest założenie o równości warunkowych stosunków szans $[XY][XZ][YZ]$. W tym przypadku, wyznaczenie rozkładu oczekiwanego wymaga zastosowania złożonych procedur iteracyjnych, niemniej również dla tej hipotezy, zgodnie z metodą największej wiarygodności, zakładamy, że:

$$\widehat{\pi}_{ij}^{XY} = p_{ij}^{XY}, \qquad \widehat{\pi}_{ik}^{XZ} = p_{ik}^{XZ} \qquad \text{oraz} \qquad \widehat{\pi}_{jk}^{YZ} = p_{jk}^{YZ}$$

Bardziej szczegółowy opis estymacji za pomocą metod iteracyjnych można znaleźć w pracy Agrestiego (1984).

### 1.3.2 Iloraz wiarygodności, statystki $G^2$ oraz $X^2$

Po oszacowaniu parametrów potrzebnych do wyznaczenia rozkładu oczekiwanego, możliwe jest — zgodnie z formułą 1.58 — wyznaczenie prawdopodobieństwa uzyskania danego wyniku w próbie losowej przy hipotezie A. Potraktujmy hipotezę $[X]$ jako hipotezę zerową $H_0$ w odniesieniu do danych z tabeli 1.23. Hipoteza konkurencyjna niech głosi $H_1 :\sim H_0$, tak więc jest równoznaczna z modelem nasyconym $[XY]$, który głosi, że rozkład dwóch zmiennych może być dowolny. Prawdopodobieństwa otrzymania danych przy każdej z dwóch rozpatrywanych hipotez $\ell_0$ oraz $\ell_1$ porównujemy za pomocą tzw. ilorazu wiarygodności:

$$\Lambda = \frac{\ell_0}{\ell_1} = \frac{P(f_{11}^{XY}, \ldots f_{ij}^{XY}, \ldots f_{rc}^{XY}|H_0)}{P(f_{11}^{XY}, \ldots f_{ij}^{XY}, \ldots f_{rc}^{XY}|H_1)} = \frac{\prod_{i=1}^{r}\prod_{j=1}^{c}(\widehat{\pi}_{ij0}^{XY})^{f_{ij}^{XY}}}{\prod_{i=1}^{r}\prod_{j=1}^{c}(\widehat{\pi}_{ij1}^{XY})^{f_{ij}^{XY}}} \tag{1.60}$$

gdzie $\widehat{\pi}_{ij0}^{XY}$ oraz $\widehat{\pi}_{ij1}^{XY}$ są prawdopodobieństwami szacowanymi związanymi odpowiednio z hipotezami $H_0$ oraz $H_1$. Prawdopodobieństwa $\widehat{\pi}_{ij0}^{XY}$ zostały oszacowane powyżej, natomiast prawdopodobieństwa dla modelu nasyconego $[XY]$ zgodnie z metodą największej wiarygodności są równe odpowiednim częstościom rozkładu łącznego z próby, tj. $\widehat{\pi}_{ij}^{XY} = p_{ij}^{XY}$. Iloraz wiarygodności — zgodnie z formułą (1.60) — jest równy w przybliżeniu 0,8.

Przyjmijmy teraz, że hipotezą zerową jest hipoteza o równomierności rozkładu łącznego $[\cdot]$. Obliczony analogicznie iloraz wiarygodności wynosi w przybliżeniu 0,166. Wielkość ta nie może być większa od 1, co wynika z tego, że hipoteza zerowa w stosunku do hipotezy alternatywnej zawiera dodatkowe założenia.

Generalnie wyższe wartości ilorazu wiarygodności przemawiają na rzecz hipotezy zerowej, niższe wartości — na rzecz hipotezy konkurencyjnej. Można jednak zadać pytanie, jakie dokładnie wartości przemawiają za odrzuceniem hipotezy zerowej, a jakie za jego za jej zaakceptowaniem? Decyzję tę podejmujemy zgodnie z regułami wnioskowania statystycznego. W tym celu konieczne jest wyznaczenie rozkładu zmiennej określającej, jakie są wartości ilorazu wiarygodności dla możliwych do wylosowania prób $n-$elementowych, przy założeniu prawdziwości hipotezy zerowej. Przypuśćmy, że rozpatrujemy hipotezę $[\cdot]$ i próby 12–elementowe. Wybrane kwantyle tak wyznaczonego rozkładu przedstawiamy w tabeli 1.24.

Tabela 1.24: Wybrane percentyle rozkładu statystyk $\Lambda$ oraz $G^2$ przy założeniu prawdziwości hipotezy $[\cdot]$

| Percentyl | $\Lambda$ | $G^2$ |
|---|---|---|
| 1–ty | 0,003 | 9,2 |
| 5–ty | 0,012 | 6,0 |
| 10–ty | 0,026 | 4,6 |
| 20–ty | 0,105 | 3,2 |
| 30–ty | 0,166 | 2,4 |
| 50–ty | 0,300 | 1,4 |
| 70–ty | 0,394 | 0,7 |
| 80–ty | 0,712 | 0,4 |

Informacje w dwóch pierwszych kolumnach tej tabeli wskazują, dla jakiego odsetka 12–elementowych prób — przy założeniu prawdziwości hipotezy zerowej $[\cdot]$— iloraz wiarygodności uzyskuje wartość nie większą niż przedstawiona w tabeli. Na przykład, jeśli w populacji rozkład łączny zmiennych jest równomierny, obliczony iloraz wiarygodności dla 5% prób będzie nie większy niż 0,012. Przypuśćmy, że przyjmiemy poziom istotności równy 0,1, tj. zakładamy, że warunkowe prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest ona prawdziwa, nie przekracza 0,1. W takiej sytuacji powinnyśmy odrzucić hipotezę zerową, gdy $\Lambda \leq 0,026$, itd. Jak widzimy, przy takim poziomie istotności, dla analizowanego przykładu nie ma potrzeby odrzucania hipotezy zerowej $[\cdot]$, gdyż wyznaczony dla próby z tabeli 1.23b iloraz wynosi 0,166.

Na ogół nie rozpatrujemy jednak rozkładu ilorazu wiarygodności, a ściśle z nim związany rozkład statystyki z próby $G^2 = -2\ln\Lambda$. Statystykę tę formułuje się zazwyczaj za pomocą liczebności z próby i liczebności oczekiwanych związanych z testowaną hipotezą. Mianowicie:

$$\begin{aligned} G^2 &= -2\ln\Lambda = 2\ln\frac{\ell_1}{\ell_0} = 2\ln\frac{\prod_{i=1}^{r}\prod_{j=1}^{c}(\widehat{\pi}_{ij1}^{XY})^{f_{ij}^{XY}}}{\prod_{i=1}^{r}\prod_{j=1}^{c}(\widehat{\pi}_{ij0}^{XY})^{f_{ij}^{XY}}} = 2\ln\prod_{i=1}^{r}\prod_{j=1}^{c}\left(\frac{\widehat{\pi}_{ij1}^{XY}}{\widehat{\pi}_{ij0}^{XY}}\right)^{f_{ij}^{XY}} = \\ &= 2\sum_{i=1}^{r}\sum_{j=1}^{c}\ln\left(\frac{\widehat{\pi}_{ij1}^{XY}}{\widehat{\pi}_{ij0}^{XY}}\right)^{f_{ij}^{XY}} = 2\sum_{i=1}^{r}\sum_{j=1}^{c}f_{ij}^{XY}\ln\left(\frac{\widehat{\pi}_{ij1}^{XY}}{\widehat{\pi}_{ij0}^{XY}}\right) = \\ &= 2\sum_{i=1}^{r}\sum_{j=1}^{c}f_{ij}^{XY}\ln\left(\frac{f_{ij}^{XY}}{\widehat{F}_{ij}^{XY}}\right) \end{aligned} \tag{1.61}$$

W przypadku rozkładu trzech lub większej liczby zmiennych formuła wyglądałaby analogicznie. Jak widać, statystyka $G^2$ zdaje sprawę z rozbieżności pomiędzy rozkładem oczekiwanym i rozkładem w próbie. Rozkład tej statystyki z próby dla hipotezy $[\cdot]$ — a dokładniej wybrane percentyle tego rozkładu — zamieściliśmy w ostatniej kolumnie tabeli 1.24. Wnioskowanie na podstawie tej statystyki można przeprowadzić analogicznie jak przedstawiliśmy to dla ilorazu wiarygodności, z tą różnicą, że wysokie jej wartości będą nas skłaniać do odrzucenia hipotezy zerowej a niskie do jej akceptacji. W praktyce, zazwyczaj posługujemy się statystyką $G^2$. Jak wykazał Wilks (1935, 1938) — w sytuacji gdy hipoteza zerowa jest prawdziwa jej rozkład dąży asymptotycznie do rozkładu $\chi^2$ o określonej liczbie stopni swobody (df), tj. nie odbiega znacząco od tego rozkładu, jeśli posługujemy się wystarczająco dużą próbą. Nie ma więc konieczności konstruowania rozkładu tej zmiennej na podstawie formuł (1.60) i (1.61) - tak jak robiliśmy to w tabeli 1.24 - co byłoby kłopotliwe obliczeniowo.

Kilka słów wyjaśnienia wymaga liczba stopni swobody związanych z poszczególnymi hipotezami. Zdaje ona sprawę z liczby niezależnych założeń, które głosi mo-

del. Przypomnijmy, że założenia dotyczyć mogą przykładowo równości odpowiednich prawdopodobieństw, szans i stosunków szans. Im prostszy jest model, tj. im więcej założeń głosi odnośnie rozkładu łącznego, tym większa jest jego liczba stopni swobody. Przykładowo, rozpatrywany jest rozkład łączny zmiennych $X$ i $Y$. Model $[X]$ głosi, że:

- Zmienne są niezależne stochastyczne, oznacza to, że rozkłady warunkowe zmiennej $X$ względem $Y$ są identyczne, tj.

$$\begin{aligned} \pi^{X(Y)}_{1(1)} &= \pi^{X(Y)}_{1(2)} = \ldots = \pi^{X(Y)}_{1(c)}, \\ \pi^{X(Y)}_{2(1)} &= \pi^{X(Y)}_{2(2)} = \ldots = \pi^{X(Y)}_{2(c)}, \\ &\ldots \\ \pi^{X(Y)}_{(r-1)(1)} &= \pi^{X(Y)}_{(r-1)(2)} = \ldots = \pi^{X(Y)}_{(r-1)(c)}. \end{aligned}$$

  Zauważmy, że w każdym wierszu mamy $(c-1)$ założeń dotyczących równości warunkowych prawdopodobieństw. Wierszy jest $(r-1)$, nie ma potrzeby wprowadzania warunku dla kategorii $x_r$, gdyż identyczność prawdopodobieństw dotyczących tej kategorii wynika z pozostałych wierszy, tj. warunek $\pi^{X(Y)}_{r(1)} = \pi^{X(Y)}_{r(2)}$ wynikałby z pozostałych warunków, nie byłby od nich niezależny. Jak widać niezależność stochastyczna $X$ i $Y$ jest definiowana przez $(r-1)(c-1)$ niezależnych warunków.

- Rozkład zmiennej $Y$ jest równomierny, czyli:

$$\pi^Y_1 = \pi^Y_2 = \ldots = \pi^Y_c,$$

  czyli odpowiednich warunków jest $(c-1)$.

Liczba stopni swobody dla modelu $[X]$ jest równa łącznej liczbie warunków określonych powyżej[7], czyli:

$$(r-1)(c-1) + (c-1) = r(c-1).$$

Przykładowo, liczba stopni swobody dla bardziej skomplikowanego modelu niezależności jest mniejsza i wynosi $(r-1)(c-1)$. Model prostszy $[\cdot]$, posiada w stosunku do

[7] Tę samą hipotezę można sformułować za pomocą szans i stosunków szans, tj. wszystkie możliwe do wyróżnienia stosunki szans są równe 1, podobnie równe 1 są wszystkie szanse związane ze zmienną Y. Jak zostanie pokazane w rozdziale drugim, wystarczy określić wartości dla szans i lokalnych stosunków szans wyróżnionych dla sąsiednich kategorii jednej i drugiej zmiennej. Na ich podstawie można określić, ile wynoszą pozostałe szanse i stosunki szans. Dlatego liczba niezależnych od siebie warunków wynosi $(r-1)(c-1)$ w odniesieniu do stosunków szans i $(c-1)$ w odniesieniu do szans, co koresponduje z liczbą warunków sformułowanych w odniesieniu do prawdopodobieństw.

modelu $[X]$ dodatkowo $(r-1)$ związanych z równomiernością rozkładu zmiennej $X$, więc liczba stopni swobody dla takiego modelu jest większa i wynosi $(rc-1)$.

Liczbę stopni swobody można też określić inaczej. Im założeń związanych z modelem jest mniej, tym więcej prawdopodobieństw trzeba szacować na podstawie próby. Liczbę stopni swobody można określić za pomocą formuły:

$$df = k - 1 - p, \tag{1.62}$$

gdzie $k$ oznacza liczbę komórek rozkładu łącznego, a $p$ liczbę niezależnych parametrów (prawdopodobieństw), które musimy oszacować na podstawie próby. Dla przykładu: w przypadku hipotezy $[\cdot]$ i rozkładu dwóch zmiennych, liczba stopni swobody wynosi: $rc-1$, gdyż nie musimy szacować na podstawie próby żadnego parametru. W przypadku hipotezy $[X]$ musimy oszacować na podstawie próby $r-1$ prawdopodobieństw związanych ze zmienną $X$, więc liczba stopni swobody dla tego modelu wynosi $rc-1-(r-1) = r(c-1)$.

Koncepcję liczby stopni swobody możemy również wytłumaczyć odwołując się parametrów modelu logarytmiczno–liniowego. Jak zostało powiedziane, aby sformułować model zgodny z daną hipotezą wybranym parametrom modelu nasyconego należy przypisać wartość 1 (lub 0 w przypadku addytywnej formy modelu). W ten sposób uzyskujemy model prostszy. Liczba stopni swobody jest określona przez liczbę niezależnych parametrów którym przypisaliśmy wartość 1 (lub odpowiednio 0). Na przykład, model $[X]$ ma jeden parametr $d$ oraz $(c-1)$ niezależnych parametrów $d_j^X$. Model nasycony ma dodatkowo $(r-1)$ niezależnych parametrów $d_i^Y$, oraz $(r-1)(c-1)$ niezależnych parametrów interakcji $d_{ij}^{XY}$. Różnica pomiędzy liczbą parametrów dla modelu nasyconego i modelu $[X]$ określa nam jego liczbę stopni swobody, tj. $df = (r-1)(c-1) + (c-1) = r(c-1)$. Tabele 1.25 i 1.26 pokazują liczbę stopni swobody dla modeli jakie można sformułować dla rozkładu dwóch i trzech zmiennych[8].

W podobny sposób do testowania modeli logarytmiczno–liniowych można stosować zaproponowaną przez Karla Pearsona statystykę $X^2$, która również porównuje liczebności oczekiwane z liczebnościami empirycznymi, tj.

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(f_{ij}^{XY} - \widehat{F}_{ij}^{XY}\right)^2}{\widehat{F}_{ij}^{XY}} \tag{1.63}$$

Analogicznie można sformułować wzór na statystykę $X^2$ dla większej liczby zmiennych. Podobnie jak statystyka $G^2$ jej rozkład dąży asymptotycznie do rozkładu $\chi^2$, je-

[8]W przypadku trzech zmiennych ograniczyliśmy się do typów modeli, dla pozostałych modeli liczbę stopni swobody można wyznaczyć analogicznie.

Tabela 1.25: Liczba parametrów i stopni swobody dla poszczególnych modeli dla dwóch zmiennych

| Oznaczenie modelu | $\pi_{ij}^{XY} = \ldots$ | Liczba niezależnych parametrów | Liczba stopni swobody |
|---|---|---|---|
| $[XY]$ | $d \cdot d_i^X \cdot d_j^Y \cdot d_{ij}^{XY}$ | rc | 0 |
| $[X][Y]$ | $d \cdot d_i^X \cdot d_j^Y$ | (r-1)+(c-1)+1=r+c-1 | (r-1)(c-1) |
| $[X]$ | $d \cdot d_i^X$ | (r-1) +1 =r | (r-1)(c-1)+(c-1)=r(c-1) |
| $[Y]$ | $d \cdot d_j^Y$ | (c-1) +1 = c | (r-1)(c-1)+(r-1)=c(r-1) |
| $[\cdot]$ | $d$ | 1 | $rc-1$ |

Tabela 1.26: Liczba parametrów i stopni swobody dla poszczególnych typów modeli dla trzech zmiennych

| Oznaczenie modelu | $\pi_{ijk}^{XYZ} = \ldots$ | Liczba niezależnych parametrów | Liczba stopni swobody |
|---|---|---|---|
| [XYZ] | $d\ d_i^X d_j^Y d_k^Z d_{ij}^{XY} d_{ik}^{XZ} d_{jk}^{YZ} d_{ijk}^{XYZ}$ | rct | 0 |
| [XY][XZ][YZ] | $d\ d_i^X d_j^Y d_k^Z d_{ij}^{XY} d_{ik}^{XZ} d_{jk}^{YZ}$ | rct-(r-1)(c-1)(t-1) | (r-1)(c-1)(t-1) |
| [XY][XZ] | $d\ d_i^X d_j^Y d_k^Z d_{ij}^{XY} d_{ik}^{XZ}$ | r(c+t-1) | r(c-1)(t-1) |
| [XY][Z] | $d\ d_i^X d_j^Y d_k^Z d_{ij}^{XY}$ | rc+t-1 | (t-1)(rc-1) |
| [XY] | $d\ d_i^X d_j^Y d_{ij}^{XY}$ | rc | rc(t-1) |
| [X][Y][Z] | $d\ d_i^X d_j^Y d_k^Z$ | r+c+t-2 | rct-r-c-t+2 |
| [X][Y] | $d\ d_i^X d_j^Y$ | r+c-1 | rct-r-c+1 |
| [X] | $d\ d_i^X$ | r | rct-r |
| $[\cdot]$ | $d$ | 1 | rct-1 |

śli hipoteza zerowa jest prawdziwa. Można zadać pytanie jak duża powinna być próba, aby przy zastosowaniu statystyk $X^2$ oraz $G^2$ było możliwe posługiwanie się rozkładem granicznym $\chi^2$. Zazwyczaj wskazuje się, że liczebność próby powinna być co najmniej dziesięciokrotnie większa od liczby komórek rozkładu łącznego (Fienberg 1980). Cochran (1954) podaje, że liczebności oczekiwane powinny być większe od 5 dla conajmniej 80% komórek rozkładu. Jeśli warunki powyższe nie są spełnione możliwe jest łączenie ze sobą poszczególnych kategorii[9]. Należy jednak pamiętać, że mamy wówczas do czynienia z hipotezą dotyczącą innych zmiennych, np. hipoteza, że płeć jest niezależna od wykształcenia mierzonego na skali 1. podstawowe, 2. średnie 3. wyższe, nie jest tożsama z hipotezą o niezależności płci i zmiennej dychotomicznej zdającej sprawę z tego, czy osoba ma wyższe wykształcenie, czy też nie. Z pierwszej

[9]Statystyczne kryteria dotyczące możliwości łączenia ze sobą kategorii zmiennej przedstawia Goodman (1981*a*).

hipotezy wynika druga, ale nie na odwrót. W przypadku, gdy mamy do czynienia z małą próbą możliwe jest również dokładne odtworzenie rozkładu statystyki $X^2$ oraz $G^2$, tak jak zrobiliśmy to w odniesieniu do hipotezy $[\cdot]$ w tabeli 1.24. Pamiętać jednak należy, że testy oparte na małej próbie cechuje na ogół mniejsza moc, tj. zwiększa się prawdopodobieństwo, że nie odrzucimy hipotezy zerowej, gdy jest ona błędna.

Posługiwanie się statystykami $X^2$ oraz $G^2$ prowadzi zazwyczaj do tego samego rezultatu, tj. jeśli odrzucamy hipotezę za pomocą statystyki $X^2$, to zwykle odrzucimy ją również stosując statystykę $G^2$. Niemniej, nawet dla bardzo dużych prób wartości tych statystyk mogą się od siebie różnić, i — co się z tym wiąże — różne mogą być wyniki weryfikacji hipotezy. Porównanie zastosowania obydwu statystyk (Agresti 2002) pokazuje, że wraz ze wzrostem liczebności próby rozkład statystyki $X^2$ szybciej dąży do rozkładu $\chi^2$ niż rozkład statystyki $G^2$. Wzór 1.63 pokazuje, że przy obliczaniu statystyki $X^2$ zakłada się jedynie, że liczebności oczekiwane powinny być większe od 0. Natomiast ze wzoru 1.61 wynika, że obliczenie statystyki $G^2$ wymaga, aby zarówno liczebności oczekiwane, jak też empiryczne były różne od 0. W takich przypadkach, na ogół dodaje się do każdej komórki niewielką stałą (np. 0,1), co umożliwia weryfikację za pomocą statystyki $G^2$, a nie zmienia w sposób istotny relacji pomiędzy liczebnościami w poszczególnych komórkach. Rozwiązanie to jest stosowane w większości pakietów statystycznych[10].

Zastosowanie statystyki $G^2$ ma pewną istotną przewagę nad statystyką $X^2$. Pozwala ona na przeprowadzanie testów warunkowych, tj. porównywanie ze sobą dwóch modeli, które różnią się ze sobą pod względem prostoty, czyli tzw. modeli *zagnieżdżonych*. Model A jest zagnieżdżony w modelu B, gdy możemy uzyskać model A z modelu B, czyniąc pewne dodatkowe założenia dotyczące np. równomierności rozkładu, zależności stochastycznej dla wybranej pary zmiennych, itd. Przekładając to na język parametrów modelu logarytmiczno–liniowego, jeśli dodatkowe założenie, że w modelu B, niektóre jego parametry są równe 1, bądź są sobie równe[11] prowadzi do sformułowania modelu A, oznacza to, że model A jest zagnieżdżony w modelu B. Na przykład: model zgodny z hipotezą o równomierności rozkładu łącznego $[\cdot]$ jest zagnieżdżony w modelu o niezależności stochastycznej $[X][Y]$, gdyż pierwszy z nich możemy uzyskać po przyjęciu dodatkowych założeń o równomierności obydwu zmiennych, tj. $d_i^X = 1$ oraz $d_j^Y = 1$ dla wszystkich wartości $x_i$ oraz $y_j$ zmiennych $X$ oraz $Y$. Podobnie model $[X][Y][Z]$ jest zagnieżdżony w modelu $[XY][Z]$, gdyż przy-

[10]Podobne założenie, przyjęte zostało w tabeli 1.24. Uwzględnione zostały wszystkie próby, jakie można wylosować, dla niektórych z nich niektóre liczebności były zerowe, wówczas dodano do wszystkich komórek stałą 0,005.

[11]Przykłady modeli zakładające równość pewnych parametrów zostaną przedstawione w dalszej części tej pracy.

jęcie dodatkowo w drugim modelu, że zmienne $X$ oraz $Y$ są niezależne stochastycznie względem $Z$ (tj. $d_{ij}^{XY} = 1$) prowadzi do sformułowania pierwszego z nich.

Okazuje się, że jeśli model B jest prawdziwy, to różnica w dopasowaniu do danych pomiędzy modelami tj. $G^2[A|B] = G^2[A] - G^2[B]$ dąży asymptotycznie do rozkładu $\chi^2$ o liczbie stopni swobody $df_{A|B} = df_A - df_B$. Wynika to z tego, że statystyka $G^2[A|B]$ jest związana z ilorazem wiarygodności porównującym prawdopodobieństwo uzyskania danych przy założeniu modelu A i modelu B, tj.

$$G^2[A|B] = G^2[A] - G^2[B] = -2\ln\frac{\ell_A}{\ell_1} - (-2\ln\frac{\ell_B}{\ell_1}) = -2\ln\frac{\ell_A}{\ell_B} \tag{1.64}$$

Określona powyżej statystyka będzie zawsze nieujemna co wynika z tego, że $G^2[A] \geq G^2[B]$, a iloraz wiarygodności porównujący model A z modelem B nie może być większy od 1. Dla przykładu, model $[\cdot]$ jest zagnieżdżony w modelu $[X]$. Dla danych z tabeli 1.23 iloraz wiarogodności wynosi $\ell_{[\cdot]}/\ell_{[X]} \approx 0,166/0,8 \approx 0,2075$, co oznacza, że uzyskanie danych z próby, takich jak w tabeli 1.23a jest prawie 5-krotnie mniej prawdopodobne, jeśli dane w populacji byłyby zgodne z modelem $[\cdot]$, niż gdyby były zgodne z modelem $[X]$.

W zwykłym teście $G^2$ model prostszy zawsze porównywany jest z modelem nasyconym. W tym sensie jest to test bezwarunkowy, gdyż model nasycony nie jest związany z żadna hipotezą. Testy warunkowe są użyteczne do odpowiedzi na pytanie, czy uproszczenie modelu pogarsza jego dopasowanie do danych w sposób istotny statystycznie. Pozwala też na przetestowanie założenia, które różni dwie porównywane hipotezy: w przypadku modeli $[\cdot]$ oraz $[X]$ testujemy hipotezę o warunkowej równomierności zmiennej $X$ względem $Y$. Hipotezę tę możemy sformułować również w języku parametrów[12] , tj. $d_i^X = 1$, lub analogicznie w wersji addytywnej $\lambda_i^X = 0$, dla każdego $i$.

Warto podkreślić zalety testów warunkowych: w porównaniu do testów bezwarunkowych mają większą one większą moc, tj. bardziej prawdopodobne jest, że odrzucimy hipotezę zerową, gdy jest ona błędna (Goodman 1981*b*). Bardziej ogólnie, można pokazać, że jeśli testujemy A przeciwko modelowi B i model A jest zagnieżdżony w modelu B, moc testu jest tym większa, im prostszy jest model B, przy założeniu jest on prawdziwy. Ponadto, jak pokazuje Haberman (1978) w sytuacji, gdy mamy do czynienia z rozkładem łącznym, w którym wiele komórek posiada bardzo małe liczebności, w przypadku testów warunkowych w porównaniu do bezwarunkowych istnieją mocniejsze podstawy by oczekiwać, że rozkład statystyki $G^2$ nie będzie znacząco odbiegał od rozkładu teoretycznego $\chi^2$. Warto w tym miejscu zaznaczyć, że statystyka $X^2$ nie

[12]Istnieją również testy, które bezpośrednio pozwalają testować hipotezy dotyczące poszczególnych parametrów. Możliwe jest również wyznaczenie przedziału ufności dla konkretnego parametru.

pozwala na porównywanie modeli zagnieżdżonych. Różnica $X^2[A] - X^2[B]$ nie dąży asymptotycznie do rozkładu $\chi^2$, co więcej może się nawet zdarzyć, że $X^2[A] < X^2[B]$.

### 1.3.3 Inne metody oceny dopasowania modelu

Testy oparte na statystykach — $X^2$ oraz $G^2$ — w przypadku posługiwania się próbą o bardzo dużej liczebności, prowadzą zazwyczaj do odrzucenia hipotezy prostszej na rzecz hipotezy bardziej złożonej. W praktyce, często okazuje się, że na konwencjonalnie przyjmowanym poziomie istotności $\alpha = 0,05$ wszystkie modele — poza nasyconym — są nieakceptowalne. Zauważmy, że formułę (1.63) na statystykę $X^2$ możemy zapisać jako:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(f_{ij}^{XY} - \widehat{F}_{ij}^{XY}\right)^2}{\widehat{F}_{ij}^{XY}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(np_{ij}^{XY} - n\widehat{\pi}_{ij}^{XY}\right)^2}{n\widehat{\pi}_{ij}^{XY}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n\left(p_{ij}^{XY} - \widehat{\pi}_{ij}^{XY}\right)^2}{\widehat{\pi}_{ij}^{XY}}$$

Jak widać statystyka ta wskazuje na rozbieżności pomiędzy częstościami z próby i odpowiednimi oszacowanymi prawdopodobieństwami zgodnymi z rozpatrywanym modelem, przy czym jest tym większa, im większa jest liczebność próby. Wynika to z tego, że jeśli rozpatrywana hipoteza jest prawdziwa, można oczekiwać, że w przypadku dużej próby rozbieżności pomiędzy rozbieżności pomiędzy rozkładem częstości w próbie i modelem będą mniejsze niż w przypadku małej próby. Podobnie jest w przypadku statystyki $G^2$. W konsekwencji, gdy liczebność próby jest bardzo duża nawet niewielkie rozbieżności pomiędzy rozkładem oczekiwanym i rozkładem z próby prowadzić mogą do odrzucenia modelu prostszego, nawet jeśli opisuje on adekwatnie badane zjawisko.

Ponieważ posługiwanie się bardzo dużą próbą prowadzi zazwyczaj do odrzucenia wszystkich modeli poza modelem nasyconym, stawia to badacza w trudnej sytuacji. Z modelem nasyconym nie jest związana żadna hipoteza, odzwierciedla on jedynie dane uzyskane w próbie losowej. Aby zdecydować, czy hipotezy prostszej nie daje się obronić, zwłaszcza, jeśli przemawiają za nią mocne argumenty teoretyczne, poza przytoczonymi powyżej statystykami $G^2$ oraz $X^2$ stosuje się inne kryteria i procedury. Wiele z nich może pomóc badaczowi przy wyborze modelu przy posługiwaniu się bardzo dużą próbą. Poniżej przedstawione zostaną alternatywne metody weryfikacji najczęściej stosowane w modelowaniu logarytmiczno–liniowym.

#### Przyjmowanie niższego poziomu istotności

Jedną ze stosowanych przez badaczy strategii jest dostosowywanie poziomu istotności do wielkości próby. Jak wiadomo, poziom istotności ($\alpha$) określa warunkowe prawdo-

podobieństwo popełnienia błędu I rodzaju, tj. odrzucenia hipotezy zerowej w sytuacji gdy jest ona prawdziwa. Przy ustalonej wielkości próby zmniejszenie poziomu istotności wiąże się na ogół ze zwiększeniem warunkowego prawdopodobieństwa popełnienia błędu II rodzaju ($\beta$), (tj. nie odrzucenia hipotezy zerowej gdy prawdziwa jest hipoteza alternatywna).

Poziom istotności jest ustalany przez badacza, więc jego wielkość nie zależy od wielkości próby. Niemniej, należy zauważyć, że wielkość warunkowego prawdopodobieństwa popełnienia błędu II rodzaju jest na ogół tym mniejsza, im większą próbą posługujemy się przy weryfikacji, mówiąc inaczej, przy większej próbie mniej jest prawdopodobne, że nie odrzucimy błędnej hipotezy zerowej, co oznacza, że moc testu jest większa [13].

Z powyższych względów, wielu badaczy zaleca, aby w przypadku posługiwania się dużą próbą zmniejszyć poziom istotności, gdyż można mieć nadzieję, że taka strategia prowadzi do zminimalizowania prawdopodobieństwa popełnienia jakiegokolwiek błędu podczas weryfikacji [14]. Przy niższym poziomie istotności bardziej jest prawdopodobne, że nie odrzucimy modelu prostego adekwatnie opisującego rzeczywistość, zauważmy jednak, że trudno określić jak niski powinien być poziom istotności przy danej wielkości próby. Sformułowanie algorytmu, w którym poziom istotności zależałby od wielkości próby, byłoby niezgodne z ideą weryfikacji hipotez Neymana-Pearsona, gdyż poziom istotności powinien wynikać z decyzji badacza.

### Indeks rozbieżności

Oceniając dopasowanie modelu do danych, można zapytać jak duży jest odsetek osób (obiektów) w przebadanej próbie, które należałoby inaczej zaklasyfikować w tabeli rozkładu łącznego, aby dane z próby idealnie odzwierciedlały model. Informuje nas o tym *indeks rozbieżności* (*dissimilarity index*). Formalnie jest on równy:

$$\Delta = \frac{\sum_{i=1}^{r} \sum_{j=1}^{c} |p_{ij}^{XY} - \widehat{\pi}_{ij}^{XY}|}{2} \tag{1.65}$$

[13] Określenie warunkowego prawdopodobieństwa popełnienia błędu II rodzaju wymaga, aby hipoteza alternatywna była hipotezą prostą, w przypadku modeli logarytmiczno–liniowych jest ona hipotezą złożoną, dlatego posługujemy się pojęciem mocy testu, które jest ogólniejsze.

[14] Wielkości $\alpha$ i $\beta$ określają warunkowe prawdopodobieństwa popełnienia błędu I i II rodzaju, natomiast prawdopodobieństwo popełnienia jakiegokolwiek błędu wymaga uwzględnienia prawdopodobieństw hipotezy zerowej i alternatywnej, tj. jest ono określone przez wyrażenie $\alpha P(h_0) + \beta P(h_1)$. Na ogół nie znamy prawdopodobieństwa tego, ze stan rzeczy jest zgodny z hipotezą zerową lub hipotezą konkurencyjną. O ile jednak $P(h_1)$ nie jest zdecydowanie większe od $P(h_0)$ można mieć nadzieję, że przy bardzo dużej liczebności i względnie małej wartości $\beta$ przyjęcie bardzo małego poziomu istotności zmniejszy prawdopodobieństwa jakiegokolwiek błędu.

W przypadku większej liczby zmiennych wzór wygląda analogicznie. Indeks rozbieżności jest równy połowie sumy absolutnych różnic pomiędzy częstościami z próby ($p_{ij}^{XY}$) oraz oszacowanymi prawdopodobieństwami oczekiwanymi ($\widehat{\pi}_{ij}^{XY}$), przy czym sumowanie dotyczy wszystkich komórek rozkładu łącznego zmiennych $X$ i $Y$. Gdybyśmy mieli do czynienia z większą liczbą zmiennych formuła byłaby analogiczna.

Indeks ten przyjmuje wartości od 0 do 1, przy czym niższe wartości wskazują na lepsze dopasowanie modelu do danych. Wartość 0 wskazuje, że rozkład w próbie całkowicie pokrywa się z rozkładem oczekiwanym. Uzyskanie wartości 1 rozważanego indeksu oznaczałoby, że wszystkie osoby w próbie należałoby przesunąć do innych komórek, aby dane z próby pokrywały się z modelem, tj. żadna osoba w próbie nie lokuje się w komórce, która zgodnie z modelem ma niezerowe prawdopodobieństwo. Należy jednak zauważyć, że dla większości modeli logarytmiczno–liniowych wartość 1 nie może być osiągnięta, gdyż nie jest możliwe skonstruowanie rozkładu całkowicie rozbieżnego z modelem. Przykładowo zgodnie z modelem niezależności stochastycznej, oszacowany rozkład oczekiwany będzie miał niezerowe prawdopodobieństwa dla każdej kombinacji wartości zmiennej wierszowej i kolumnowej, która pojawia się w próbie. Dla tego modelu nie jest więc możliwe osiągnięcie wartości indeksu rozbieżności równej 1.

Indeks rozbieżności początkowo wykorzystywany był do opisu rozbieżności pomiędzy rozkładem zmiennej w dwóch populacjach. Goodman i Kruskal (1959) wskazują, że jako pierwszy zaproponował go Gini (1914). Zauważmy, że miara ta zdaje jedynie sprawę, jaki odsetek osób w jednej populacji powinien być „przesunięty" aby rozkłady w obydwu populacjach były zgodne, przy czym nie bierze pod uwagę typu „przesunięć". Wydaje się to niepożądana własność na przykład w odniesieniu do zmiennej porządkowej, która opisuje opinie respondenta na skali, czym innym jest przesunięcie od odpowiedzi „zdecydowanie się zgadzam" do odpowiedzi „raczej się zgadzam" niż do odpowiedzi „zdecydowanie się nie zgadzam".

Indeks rozbieżności, choć bywa pomocny do oceny modelu, nie powinien być jedynym kryterium. Jedną z jego wad jest trudność w określeniu wartości, od której należałoby akceptować dany model. Trudno powiedzieć, czy dana wartość jest wystarczająco duża by model odrzucić. Ponadto, jak pokazuje formuła 1.65 indeks nie zależy od wielkości próby. Przypuśćmy, że oceniamy ten sam model na podstawie próby 100 i 1000–elementowej i uzyskujemy taką samą wartość indeksu rozbieżności równą 0,05. Wydaje się, ze w drugim przypadku jest to poważniejsza przesłanka do odrzucenia modelu: można bowiem przypuszczać, że jeśli model jest prawdziwy to częstości z większej próby powinny być bardziej zbliżone do tego co dzieje się w populacji.

Wady posługiwaniem się indeksem rozbieżności daje się w pewnej mierze wyeliminować[15]. Możliwe jest wyznaczenie rozkładu indeksu rozbieżności jako statystyki z próby i przeprowadzenie weryfikacji podobnie jak w przypadku statystyk $G^2$ oraz $X^2$. W pracy tej strategia ta nie będzie stosowana: indeks rozbieżności będzie wykorzystywany jedynie jako miara dodatkowa, która może ustrzec przed pochopnym odrzuceniem modelu. Przykładowo, jeśli badacz posługuje się bardzo dużą próbą, statystyki $G^2$ oraz $X^2$ pokazują, że na konwencjonalnie przyjmowanym poziomie istotności 0,01 lub 0,05 powinniśmy model odrzucić, a indeks rozbieżności pokazuje, że mniej niż 1% osób w próbie jest zaklasyfikowanych niezgodnie z modelem, może skłonić to badacza do zaakceptowania modelu jako adekwatnego opisu rzeczywistości, zwłaszcza jeśli model posiada uzasadnienie teoretyczne.

### Bayesowskie podejście do weryfikacji hipotez. Indeks BIC

Alternatywne podejście do wnioskowania statystycznego związane jest z wykorzystaniem twierdzenia Bayesa. W modelowaniu logarytmiczno–liniowym szczególnie często stosowany jest indeks BIC (Bayesian Information Criterion) wprowadzony przez Raftery'ego (1986*a*, 1986*b*, 1995). Jest on jedną z możliwych implementacji tego podejścia. Indeks ten definiuje się następująco:

$$BIC = G^2 - (\ln n)df \tag{1.66}$$

Z jednej strony bierze on pod uwagę stopień rozbieżności między rozkładem oczekiwanym i rozkładem z próby ($G^2$), z drugiej strony, prostotę modelu, określoną przez liczbę stopni swobody (df) i wielkość próby (a dokładniej jej logarytm). Dla modelu nasyconego wartość tego indeksu jest równa 0, wartości mniejsze od 0 wskazują na modele akceptowalne. Zaletą indeksu BIC, jest to, że pozwala on na porównywanie ze sobą dopasowania dwóch, niekoniecznie zagnieżdżonych modeli (inaczej niż w przypadku statystyki $G^2$). Porównując dwa modele, wybieramy ten o niższej wartości BIC.

Weryfikacja hipotez w ujęciu bayesowskim została zapoczątkowana przez Jeffrey'a w latach 30-tych (Jeffreys 1961, Raftery i Kass 1995). Zaproponowana idea dotyczyła porównania mocy predyktywnej dwóch konkurujących ze sobą teorii. Zgodnie z tym ujęciem modele statystyczne powiązane z tymi teoriami, powinny być konstruowane w taki sposób, aby można było oszacować prawdopodobieństwo otrzymania wyników testu empirycznego (np. danych z próby losowej) przy założeniu prawdziwości każdej teorii. W omawianym podejściu prawdopodobieństwa te wykorzystywane są do

[15]Kuha i Firth (2004) omawiają szczegółowo, również inne wady tej miary, jak również przedstawiają propozycje modyfikacji indeksu.

policzenia tzw. prawdopodobieństw *a posteriori*. Pojęcie to opisuje prawdopodobieństwo tego, że dany model $M_1$ jest prawdziwy, przy założeniu otrzymaniu konkretnych danych z próby. Będziemy je oznaczać jako $P(M_1|D)$. W przypadku, gdy rozważamy dwa konkurujące ze sobą modele $M_1$ i $M_2$, prawdopodobieństwo *a posteriori* dla modelu $M_1$ można przedstawić wykorzystując twierdzenie Bayesa:

$$P(M_1|D) = \frac{P(D|M_1)P(M_1)}{P(D)} = \frac{P(D|M_1)P(M_1)}{P(D|M_1)P(M_1) + P(D|M_2)P(M_2)} \tag{1.67}$$

gdzie $P(D|M_1)$ określa prawdopodobieństwo uzyskania danych z próby przy założeniu, że prawdziwy jest model $M_1$, a $P(M_1)$ jest określonym *a priori* prawdopodobieństwem, że prawdziwy jest model $M_1$. Porównując ze sobą prawdopodobieństwa *a posteriori* dla dwóch modeli $M_1$ i $M_2$, definiujemy, tzw. szansę *a posteriori* (*posterior odd*):

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)P(M_1)}{P(D|M_2)P(M_2)} \tag{1.68}$$

Modele $M_1$ i $M_2$ mogą być dowolnymi modelami, niekoniecznie jeden z nich musi być zagnieżdżony w drugim. Badacz na ogół nie ma wiedzy, odnośnie tego, który z modeli $M_1$, $M_2$ jest bardziej prawdopodobny. Czasem zakłada się arbitralnie, że: $P(M_1) = P(M_2)$, aczkolwiek należy pamiętać, że założenie to może nie być trafne. Zdefiniowana powyżej szansa *a posteriori* przy takim założeniu sprowadza się wówczas do:

$$B_{12} = \frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)} \tag{1.69}$$

Wielkość $B_{12}$ określa, czy przy obserwowanych danych bardziej prawdopodobny jest model $M_1$ czy $M_2$ i nazywa się ją *czynnikiem bayesowskim* (*bayes factor*). Prawdopodobieństwo $P(D|M_1)$ jest określone przez:

$$P(D|M_1) = \int P(D|\epsilon_\mathbf{1})P(\epsilon_\mathbf{1}|M_1)d\epsilon_1 \tag{1.70}$$

gdzie $\epsilon_\mathbf{1}$ jest parametrem, lub wektorem parametrów określonych przez model $M_1$. Przypuśćmy, że rozważamy rozkład dwóch dychotomicznych zmiennych $X$ i $Y$. Porównujemy ze sobą model niezależności stochastycznej i model nasycony. Różnią się one występowaniem parametru interakcji $\lambda_{11}^{XY}$ (lub jego multiplikatywnym odpowiednikiem $d_{11}^{XY}$)[16]. Wielkość ta wskazuje na siłę związku między zmiennymi. W modelu niezależności zakładamy, że parametr ten jest równy 0 (w wersji multiplikatywnej 1). W odniesieniu do modelu nasyconego badacz powinien założyć, jak jego zdaniem

[16]Równoważnie jako parametr możemy potraktować stosunek szans $\Theta_{11}^{XY}$, lub jej logarytm.

prawdopodobne są różne wartości tego parametru, przy założeniu prawdziwości modelu $M_1$, tj. $P(\epsilon_\mathbf{1}|M_1)$. Określone w ten sposób oczekiwania badacza określają tzw. rozkład *a priori* (*prior distribution*) parametru lub parametrów $\epsilon_\mathbf{1}$, związanych z hipotezą $M_1$.

Zdefiniowanie rozkładu *a priori* parametru, a szczególnie rozkładu łącznego kilku parametrów $\epsilon_\mathbf{1}$ nie zadaniem łatwym. Skomplikowane pod względem obliczeniowym jest również wykorzystanie tego rozkładu do wyznaczenia czynnika bayesowskiego. Omawiany w tym miejscu indeks BIC, zdefiniowany w formule (1.66), przy pewnych założeniach, których nie będziemy omawiać szczegółowo może być interpretowany jako czynnik bayesowski. Po uwzględnieniu tych założeń jeśli porównujemy dowolny model z modelem nasyconym, to dla wystarczająco dużej próby zachodzi:

$$BIC \approx -2 \log B_{01} \tag{1.71}$$

W tym przypadku model $M_0$ porównujemy z modelem nasyconym $M_1$. Podobnie różnica indeksu BIC dla dwóch dowolnych modeli koresponduje z czynnikiem bayesowskim dla tych modeli.

$$BIC(M_1) - BIC(M_2) \approx -2 \log B_{12} \tag{1.72}$$

Wielu badaczy przekonuje, że indeks ten w praktyce jest dobrym miernikiem oceny dopasowania modelu do danych i okazuje się szczególnie przydatny w sytuacji, gdy posługujemy się bardzo dużą próbą (Raftery 1986a, 19986b, 1995, 1999, Raftery i Kass 1995, Hagenaars 1990, Xie 1999). Indeks BIC może być traktowany jako przybliżenie czynnika bayesowskiego[17], po przyjęciu pewnych założeń dotyczących rozkładu *a priori* $P(\epsilon_\mathbf{1}|M_1)$, dzięki czemu badacz nie musi specyfikować ich samodzielnie.

Wielu innych badaczy krytykuje jednak przyjęte założenia i indeks BIC (Gelman i Rubin 1995, 1999, Weakliem 1999a, 1999b). Wskazują, że wadą indeksu BIC, jest to, że rozkład a priori nie wynika z przekonań badacza, a założenia są przyjmowane „automatycznie" na podstawie danych empirycznych, co kłóci się z ideą „baeysowkiej" weryfikacji hipotez. Założenia te mogą być nietrafne w stosunku do wielu hipotez. Wskazywano na liczne paradoksalne i niepożądane własności związane z zastosowaniem tego kryterium, ponadto nie jest prawdą, że indeks ten jest „odporny" na odrzucanie hipotezy prostszej na rzecz bardziej skomplikowanej w przypadku bardzo dużych prób. Na przykład, jak pokazuje Weakliem (2004) o ile w odniesieniu do prób liczących około 10 tysięcy obserwacji, indeks BIC często prowadzi do wyboru modelu prostszego, nawet gdy jest on odrzucony przy posługiwaniu się statystyką $G^2$ lub $X^2$, to jeśli próba liczy około 100 tysięcy obserwacji obydwie metody prowadzą zazwyczaj do wyboru modelu nasyconego.

[17] Czytelnik może znaleźć odpowiednie przekształcenia w: Raftery 1995, Raftery i Kass 1995.

### Inne miary dopasowania

W literaturze zaproponowano szereg innych miar, pomocnych do oceny modelu. Często stosowaną strategią jest odniesienie statystyki $G^2$ lub $X^2$ do wielkości próby $n$, bądź do liczby stopni swobody modelu $df$ (Bonett i Bentler 1971, Goodman 1971, 1975, Hagenaars 1990). Pierwsza z tych propozycji ($G^2/n$) ma na celu uwzględnienie faktu, że liczebność próby wpływa na wielkość tych statystyk, druga propozycja ($G^2/df$) uwzględnia, że w odniesieniu do modeli prostszych możemy oczekiwać, że statystyka $G^2$ (i podobnie $X^2$) będzie większa, gdy model jest prostszy. Zaleca się odrzucenie modelu gdy wartość pierwszej z tych miar przekracza 0,1 a drugiej znacznie przekracza 1[18] .

Goodman (1972$a$, 1972$b$) zaproponował miarę przydatną do porównywania modeli zagnieżdżonych. Miernik ten określa w jakim stopniu dodanie kolejnych parametrów do modelu poprawia dopasowanie modelu do danych. Formalnie:

$$R' = \frac{G^2(A) - G^2(B)}{G^2(A)} \tag{1.73}$$

przy czym model A jest prostszy niż model B. Jeśli $R' = 0$, oznacza to, że dopasowanie obydwu modeli jest identyczne. Przyjmuje się, że wartości wyższe niż 0,8 powinny nas skłaniać do odrzucenia modelu prostszego (Zahn i Fein 1979)[19]. Badacze często porównują wszystkie rozważane modele z jednym modelem A. Może to być szczególnie sensowne jeśli posiada on dogodną interpretację jako model „odniesienia". Na przykład, model głoszący niezależność stochastyczną trzech zmiennych, może stanowić odniesienie dla modeli opisujących różne typy zależności. To, czy któryś z modeli ma dogodną interpretację jako model odniesienia zależy od badanego zjawiska. Miara ta może być również użyteczna przy posługiwaniu się dużą próbą. Na przykład: jeśli model B nie może być zaakceptowany, w oparciu o test bezwarunkowy za pomocą statystyki $G^2$, to jej znaczna redukcja w stosunku do modelu A, może powstrzymać badacza od jego odrzucenia.

Bentler i Bonnet (1983) zaproponowali modyfikację miernika $R'$ polegającą na uwzględnieniu prostoty modelu, tj. jego liczby stopni swobody:

$$R'' = \frac{G^2(A)/df_A - G^2(B)/df_B}{G^2(A)/df_A} \tag{1.74}$$

[18]W przypadku miary $G^2/n$, reguła ta opiera się na obserwacjach empirycznych i nie ma teoretycznego uzasadnienia. Jeśli chodzi o miarę ($G^2/df$), to o ile model jest prawdziwy, statystyka ta ma rozkład, który asymptotycznie dąży do rozkładu F, przy liczbie stopni swobody $df_1 = df$, $df_2 = \infty$, gdzie $df$ wskazuje na liczbę stopni swobody rozpatrywanego modelu (Goodman 1971, 1975). Dokładniejsze omówienie własności tych statystyk przekracza jednak ramy tej pracy.

[19]Nie można więc wykorzystywać tego miernika do przeprowadzania testów bezwarunkowych, gdyż jeśli model B, jest modelem nasyconym, wartość powyższego indeksu będzie równa 1 (ponieważ statystyka dopasowania $G^2$ dla modelu nasyconego wynosi 0)

Podobnie jak w przypadku $R'$ wysokie wartości tego miernika skłaniać będą do odrzucenia modelu prostszego. Należy zauważyć, że wartości miernika $R''$ mogą być również negatywne. Dzieje się tak w sytuacji, gdy nałożone ograniczenia na model prostszy nie powodują znacznego pogorszenia dopasowania modelu do danych.

### Istotność statystyczna a wielkość efektu

Powyżej przedstawione zostały miary opisowe pozwalające ocenić dopasowanie modelu do danych. Wiele z nich może być pomocne gdy posługujemy się bardzo duża próbą, czyli w sytuacji, gdy statystyki $G^2$ lub $X^2$ na ogół prowadzą do odrzucenia hipotezy prostszej na rzecz modelu nasyconego. Nie oznacza to, że wielkość próby, nie ma wpływu na wartości tych miar ani na decyzje, które za ich pomocą się podejmuje.

Wielu badaczy zwraca uwagę na bardziej ogólny problem. W procesie weryfikacji hipotez odpowiadamy na pytanie: czy model opisuje populację? a nie na pytanie: czy model *w przybliżeniu* opisuje populację? Na przykład: testujemy hipotezę o niezależności dwóch zmiennych. Trudno oczekiwać, że zmienne w populacji są rzeczywiście idealnie niezależne. Jeśli związek ten istnieje i nawet jeśli jest bardzo słaby, wówczas bardzo duża próba na ogół prowadzi do odrzucenia hipotezy o niezależności. Zauważmy, że z perspektywy weryfikacji hipotez jest to poprawna konkluzja. Oczywiście, siła tego związku może być tak mała, że model o niezależności wydaje się z merytorycznego punktu widzenia dobrym opisem. Należy jednak podkreślić, że wskazują na to wielkości poszczególnych parametrów, natomiast przedstawione testy oparte statystykach $G^2$, $X^2$ i wiele innych miar dopasowania modelu do danych odpowiadają na pytanie czy zaobserwowane rozbieżności pomiędzy modelem a próbą wynikają jedynie z błędu losowego, nie zaś, czy rozbieżności pomiędzy modelem a populacją są tak małe, że można je pominąć. Warto również zauważyć, że przedstawiona procedura weryfikacji hipotez uwzględnia jedynie błąd losowy nie uwzględnia natomiast, że rozbieżności pomiędzy próbą losową a modelem mogą wynikać z innych rodzajów błędów, np: pomiaru, proceduralnych (pomyłek ankietera, osoby wprowadzającej dane do komputera, itd). Nawet jeśli występują one relatywnie rzadko, szczególnie przy bardzo dużej próbie, mogą one mieć wpływ na odrzucenie hipotezy zerowej. Przedstawiona procedura weryfikacji, zakładała, ze błędy te były nieobecne[20].

.

[20]Błędy te uwzględnia się czasem na etapie budowy modelu, przedstawienie tego zagadnienia przekracza jednak ramy tej pracy.

## 1.4 Przykład analizy empirycznej

W tabeli 1.27 przedstawiony został rozkład łączny trzech zmiennych: płci ($P$), przynależności społeczno–zawodowej, (klasyfikacja SKZ[21]) ($Z$), oraz deklaracji respondenta dotyczącej udziału w wyborach parlamentarnych w roku 2001[22], (zmienna $V$). Dane pochodzą z polskiej edycji I rundy międzynarodowego badania *European Social Survey*[23] (Sztabiński i Sztabiński, 2003).

Wyniki weryfikacji omówionych w tym rozdziale modeli logarytmiczno–liniowych dla trzech zmiennych zostały zamieszczone w tabeli 1.28. Przedstawione zostały wartości statystyk $\chi^2$ oraz $G^2$. W nawiasach zamieszczone zostały minimalne poziomy istotności (oznaczane na ogół jako $p$), które pozwalałyby testowany model odrzucić. Jeśli przyjęty przez nas poziom istotności ($\alpha$) jest większy od uzyskanej wielkości $p$, oznacza to, że powinniśmy testowaną hipotezę odrzucić.

Posługując się statystykami $\chi^2$ oraz $G^2$ na poziomie istotności $\alpha = 0,01$, większość modeli należałoby odrzucić, akceptowalne są jedynie modele $[PZ][VZ]$ oraz $[PV][PZ][VZ]$. Pierwszy z nich głosi, ze płeć ($P$) i udział w wyborach ($V$) są warunkowo niezależne względem przynależności społeczno–zawodowej. Model ten jest akceptowalny na poziomie istotności równym $\alpha = 0,01$, ale musiałby zostać odrzucony poziomie istotności równym $\alpha = 0,05$. Zgodnie z drugim modelem każda para zmiennych może być warunkowo zależna stochastycznie, ale siła tej zależności mierzona stosunkiem szans jest taka sama dla każdej wartości trzeciej zmiennej. Na przykład: płeć i udział w wyborach są zależne stochastycznie, tj. stosunek szans opisujący tę zależność nie musi być równy 1, ale jego wartość jest taka sama dla każdej kategorii zawodowej. To samo można powiedzieć o zależności pomiędzy płcią i przynależnością zawodową względem uczestnictwa w wyborach i związku pomiędzy zawodem i uczestnictwem w wyborach względem płci. Model ten ma 6 stopni swobody a statystyki dla

[21] Społeczna Klasyfikacja Zawodów jest rozwijana w Polsce od lat 70-tych (Pohoski, Słomczyński i Milczarek 1974, Pohoski i Słomczyński 1978). W odniesieniu do powyższych danych wykorzystana została wersja z 1995 roku (Domański i Sawiński 1995). Ostatnia modyfikacja dokonana została w 2007 roku (Domański, Słomczyński i Sawiński 2008). W przypadku tej analizy, informacja o zawodzie dotyczy pracy aktualnie bądź ostatnio wykonywanej.

[22] Badanie odbyło się rok po wyborach. W analizie uwzględnione zostały jedynie te osoby, dla których mieliśmy informację o wszystkich trzech zmiennych, czyli około 83% zrealizowanej próby.

[23] Zastosowano złożony schemat doboru próby, przy wyznaczaniu rozkładu łącznego zastosowano odpowiednią wagę, dlatego liczebności nie są liczbami całkowitymi. Prezentacja analiz ma głównie cel ilustracyjny, dlatego weryfikacja hipotez została przeprowadzona przy założeniu, że mamy do czynienia z doborem prostym zależnym. Podobna strategia — dotycząca ważenia i weryfikacji hipotez — zostanie przyjęta również w kolejnych analizach empirycznych. Dodatkowe informacje o badaniu, z którego pochodzą dane można znaleźć w Aneksie i na stronie internetowej *www.europeansocialsurvey.org*.

Tabela 1.27: Udział w wyborach parlamentarnych 2001 roku w zależności od płci i przynależności społeczno–zawodowej[a]

| Mężczyźni ($P = 1$) | | |
|---|---|---|
| Kategorie społ.–zawodowe ($Z$) | Nie głosował ($V = 1$) | Głosował ($V = 2$) |
| 1. Inteligencja | 19,7 | 85,1 |
| 2. Pozostali pracownicy umysłowi | 18,0 | 104,4 |
| 3. Pracownicy fizyczno–umysłowi | 17,0 | 50,1 |
| 4. Robotnicy wykwalifikowani | 133,1 | 243,9 |
| 5. Robotnicy nie wykwalifikowani | 21,9 | 29,7 |
| 6. Rolnicy | 20,5 | 71,5 |
| 7. Prywatni przedsiębiorcy | 11,9 | 47,9 |
| Kobiety ($P = 2$) | | |
| Kategorie społ.–zawodowe ($Z$) | Nie głosował ($V = 1$) | Głosował ($V = 2$) |
| 1. Inteligencja | 12,0 | 41,3 |
| 2. Pozostali pracownicy umysłowi | 76,9 | 217,6 |
| 3. Pracownicy fizyczno–umysłowi | 51,0 | 98,9 |
| 4. Robotnicy wykwalifikowani | 41,7 | 60,5 |
| 5. Robotnicy nie wykwalifikowani | 60,1 | 69,8 |
| 6. Rolnicy | 45,4 | 83,5 |
| 7. Prywatni przedsiębiorcy | 9,4 | 19,4 |

[a]Źródło: Europejski Sondaż Społeczny, 2002. Dane przeważone.

tego modelu wskazują na jego dobre dopasowanie do danych: $\chi^2 = 3,34$ ($p = 0,7656$) oraz $G^2 = 3,38$ ($p = 0,7602$).

Model $[PZ][VZ]$ jest modelem zagnieżdżonym w modelu $[PV][PZ][VZ]$, tj. model ten zakłada dodatkowo, że zmienne pleć i udział w wyborach są warunkowo niezależne stochastycznie. Modele te można więc porównać za pomocą warunkowego testu, jego liczba stopni swobody wynosi 1. Różnica w dopasowaniu dla obydwu modeli wynosi $G^2 = 16,08 - 3,38 = 12,7$ (p=0,0004). Oznacza to, że nawet przy bardzo niskim poziomie istotności $\alpha = 0,001$ należy odrzucić hipotezę zgodnie z którą parametr interakcji drugiego rzędu $d_{11}^{PV} = 1$. Test warunkowy wskazuje więc na model $[PV][PZ][VZ]$. Do podobnych wniosków prowadzi porównanie mierników BIC (dla tego modelu wartość tego miernika jest najniższa). Wartość indeksu rozbieżności wskazuje, że jedynie 1,6% próby jest zaklasyfikowanych niezgodnie z tym modelem.

Tabela 1.28: Wyniki weryfikacji modeli dla danych z tabeli 1.27

| Model | df | $\chi^2$ | $G^2$ | $\Delta$ | BIC |
|---|---|---|---|---|---|
| $[PV][PZ][VZ]$ | 6 | 3,3 $(p = 0,7656)$ | 3,4 $(p = 0,7602)$ | 1,6 | -41,5 |
| $[PV][PZ]$ | 12 | 56,6 $(p < 0,0001)$ | 57,5 $(p < 0,0001)$ | 7,2 | -32,2 |
| $[PV][VZ]$ | 12 | 335,9 $(p < 0,0001)$ | 351,7 $(p < 0,0001)$ | 20,6 | 262,0 |
| $[PZ][VZ]$ | 7 | 15,6 $(p = 0,0289)$ | 16,1 $(p = 0,0244)$ | 3,5 | -36,2 |
| $[PV][Z]$ | 18 | 390,4 $(p < 0,0001)$ | 399,9 $(p < 0,0001)$ | 20,7 | 265,4 |
| $[PZ][V]$ | 13 | 62,5 $(p < 0,0001)$ | 64,3 $(p < 0,0001)$ | 7,56 | -32,8 |
| $[VZ][P]$ | 13 | 341,2 $(p < 0,0001)$ | 358,5 $(p < 0,0001)$ | 20,6 | 261,3 |
| $[PV]$ | 24 | 1054,0 $(p < 0,0001)$ | 864,3 $(p < 0,0001)$ | 27,1 | 685,0 |
| $[PZ]$ | 14 | 319,4 $(p < 0,0001)$ | 337,7 $(p < 0,0001)$ | 19,4 | 233,1 |
| $[VZ]$ | 14 | 341,3 $(p < 0,0001)$ | 358,6 $(p < 0,0001)$ | 20,6 | 254,0 |
| $[P][V][Z]$ | 19 | 392,4 $(p < 0,0001)$ | 406,7 $(p < 0,0001)$ | 20,9 | 264,7 |
| $[P][V]$ | 25 | 1040,0 $(p < 0,0001)$ | 871,2 $(p < 0,0001)$ | 27,9 | 684,3 |
| $[P][Z]$ | 20 | 677,8 $(p < 0,0001)$ | 680,1 $(p < 0,0001)$ | 25,7 | 530,7 |
| $[V][Z]$ | 20 | 392,4 $(p < 0,0001)$ | 406,8 $(p < 0,0001)$ | 21,0 | 257,3 |
| $[P]$ | 26 | 1402,9 $(p < 0,0001)$ | 1144,6 $(p < 0,0001)$ | 31,5 | 950,3 |
| $[V]$ | 26 | 1038,0 $(p < 0,0001)$ | 871,3 $(p < 0,0001)$ | 27,9 | 676,9 |
| $[Z]$ | 21 | 677,4 $(p < 0,0001)$ | 680,2 $(p < 0,0001)$ | 25,7 | 523,3 |
| $[\cdot]$ | 27 | 1400,7 $(p < 0,0001)$ | 1144,7 $(p < 0,0001)$ | 31,5 | 942,9 |

W tabeli 1.29 prezentujemy parametry tego modelu. Efekty główne zmiennej $V$ wskazują, że — zgodnie z tym modelem — średnia geometryczna szans wyznaczonych dla wszystkich podzbiorowości wyróżnionych ze względu na płeć i przynależność społeczno–zawodową na spotkanie osoby głosującej w stosunku do niegłosującej[24] wynosi:

$$\Omega^{V(PZ)}_{2/1(**)} = \tau^V_2 \backslash \tau^V_1 = 1,56/0,64 = 2,43.$$

Określona w ten sposób proporcja osób głosujących do niegłosujących, jest większa wśród mężczyzn niż wśród kobiet o czym informują parametry drugiego rzędu: średnia geometryczna określonej powyżej szansy obliczona dla wszystkich kategorii

[24]Należy podkreślić że mamy do czynienia z deklaracją głosowania, co nie musi być zbieżne z faktycznym uczestnictwem respondenta w wyborach. Frekwencja wyborcza w 2001 roku nie przekraczała 50%. Rozbieżności wynikać mogą zarówno ze słabej pamięci respondentów, niechęci do przyznania się do nie–uczestniczenia w wyborach, jak również z tego, że można oczekiwać, że osoby odmawiające udziału w badaniu to relatywnie częściej osoby niegłosujące

Tabela 1.29: Parametry modelu $[PV][PZ][VZ]$

| Parametryzacja odchyleń multiplikatywnych | | | Parametryzacja względem kategorii odniesienia | | |
|---|---|---|---|---|---|
| $\tau_1^Z = 0,68$ | $\tau_{11}^{ZP} = 1,40$ | $\tau_{11}^{ZV} = 0,80$ | $\gamma_1^Z = 1,00$ | $\gamma_{11}^{ZP} = 1,00$ | $\gamma_{11}^{ZV} = 1,00$ |
| $\tau_2^Z = 1,72$ | $\tau_{12}^{ZP} = 0,71$ | $\tau_{12}^{ZV} = 1,24$ | $\gamma_2^Z = 1,18$ | $\gamma_{12}^{ZP} = 1,00$ | $\gamma_{12}^{ZV} = 1,00$ |
| $\tau_3^Z = 1,02$ | $\tau_{21}^{ZP} = 0,64$ | $\tau_{21}^{ZV} = 0,81$ | $\gamma_3^Z = 0,92$ | $\gamma_{21}^{ZP} = 1,00$ | $\gamma_{21}^{ZV} = 1,00$ |
| $\tau_4^Z = 2,13$ | $\tau_{22}^{ZP} = 1,55$ | $\tau_{22}^{ZV} = 1,24$ | $\gamma_4^Z = 6,98$ | $\gamma_{22}^{ZP} = 4,73$ | $\gamma_{22}^{ZV} = 0,99$ |
| $\tau_5^Z = 0,90$ | $\tau_{31}^{ZP} = 0,68$ | $\tau_{31}^{ZV} = 1,01$ | $\gamma_5^Z = 1,05$ | $\gamma_{31}^{ZP} = 1,00$ | $\gamma_{31}^{ZV} = 1,00$ |
| $\tau_6^Z = 1,10$ | $\tau_{32}^{ZP} = 1,47$ | $\tau_{32}^{ZV} = 0,99$ | $\gamma_6^Z = 1,23$ | $\gamma_{32}^{ZP} = 4,24$ | $\gamma_{32}^{ZV} = 0,63$ |
| $\tau_7^Z = 0,40$ | $\tau_{41}^{ZP} = 1,99$ | $\tau_{41}^{ZV} = 1,25$ | $\gamma_7^Z = 0,69$ | $\gamma_{41}^{ZP} = 1,00$ | $\gamma_{41}^{ZV} = 1,00$ |
| | $\tau_{42}^{ZP} = 0,50$ | $\tau_{42}^{ZV} = 0,80$ | | $\gamma_{42}^{ZP} = 0,49$ | $\gamma_{42}^{ZV} = 0,41$ |
| $\tau_1^P = 0,94$ | $\tau_{51}^{ZP} = 0,66$ | $\tau_{51}^{ZV} = 1,35$ | $\gamma_1^P = 1,00$ | $\gamma_{51}^{ZP} = 1,00$ | $\gamma_{51}^{ZV} = 1,00$ |
| $\tau_2^P = 1,06$ | $\tau_{52}^{ZP} = 1,52$ | $\tau_{52}^{ZV} = 0,74$ | $\gamma_2^P = 0,71$ | $\gamma_{52}^{ZP} = 4,52$ | $\gamma_{52}^{ZV} = 0,35$ |
| | $\tau_{61}^{ZP} = 0,86$ | $\tau_{61}^{ZV} = 1,00$ | | $\gamma_{61}^{ZP} = 1,00$ | $\gamma_{61}^{ZV} = 1,00$ |
| $\tau_1^V = 0,64$ | $\tau_{62}^{ZP} = 1,17$ | $\tau_{62}^{ZV} = 1,00$ | $\gamma_1^V = 1,00$ | $\gamma_{62}^{ZP} = 2,66$ | $\gamma_{62}^{ZV} = 0,65$ |
| $\tau_2^V = 1,56$ | $\tau_{71}^{ZP} = 1,45$ | $\tau_{71}^{ZV} = 0,91$ | $\gamma_2^V = 4,65$ | $\gamma_{71}^{ZP} = 1,00$ | $\gamma_{71}^{ZV} = 1,00$ |
| | $\tau_{72}^{ZP} = 0,69$ | $\tau_{72}^{ZV} = 1,10$ | | $\gamma_{72}^{ZP} = 0,93$ | $\gamma_{72}^{ZV} = 0,79$ |
| $\tau_{11}^{PV} = 0,90$ | | | $\gamma_{11}^{PV} = 1,00$ | | |
| $\tau_{12}^{PV} = 1,11$ | | | $\gamma_{12}^{PV} = 1,00$ | | |
| $\tau_{21}^{PV} = 1,11$ | | | $\gamma_{21}^{PV} = 1,00$ | | |
| $\tau_{22}^{PV} = 0,90$ | | | $\gamma_{22}^{PV} = 0,66$ | | |

zawodowych wśród mężczyzn wynosi:

$$\Omega_{2/1(1*)}^{V(PZ)} = \Omega_{2/1(**)}^{V(PZ)} \cdot \frac{\tau_{12}^{PV}}{\tau_{11}^{PV}} = 2,43 \cdot 1,11^2 = 2,99$$

analogiczna średnia dla kobiet wynosi 1,97. Podobnie można wyznaczyć średnie geometryczne szans związane z głosowaniem dla poszczególnych kategorii zawodowych (średnie te obliczamy dla szans wyznaczonych dla kobiet i mężczyzn ). Średnia ta jest najwyższa dla *inteligencji* $(2,43 \cdot 1,24^2 = 3,73)$ i *pozostałych pracowników umysłowych*, natomiast najniższa dla *robotników niewykwalifikowanych* $(2,43 \cdot 0,74^2 = 0,84)$.

Parametry interakcji pozwalają nam również zrekonstruować wartości stosunków szans. Choć możemy je wyznaczyć na podstawie parametryzacji odchyleń multiplikatywnych, wygodniej będzie posłużyć się parametryzacją względem pierwszych kategorii trzech zmiennych. Stosunek szans dla płci i uczestnictwa w wyborach wynosi $\Theta_{2/1;2/1(1)}^{P\ \ V\ (Z)} = \gamma_{22}^{PV} = 0,64$, co oznacza, że proporcja osób głosujących do niegłosujących, jest wśród kobiet ponad 1,5 razy niższa (tj. 1/0,64) aniżeli wśród mężczyzn.

Parametr $\gamma_{22}^{PV}$ opisuje stosunek szans $\Theta_{2/1;2/1(1)}^{P \quad V \ (Z)}$ w podzbiorowości odniesienia $Z = 1$, jednak zgodnie z tym modelem stosunki szans są takie same w każdej podzbiorowości wyodrębnionej ze względu na przynależność społeczno–zawodową.

# Rozdział 2

# Modele logarytmiczno–liniowe dla zmiennych porządkowych

W poprzednim rozdziale przedstawione zostały modele dla zmiennych nominalnych. Modele te można również stosować do analizy zmiennych mierzonych na mocniejszych skalach, na przykład do zmiennych porządkowych. Należy jednak zauważyć, że nie jest w nich wykorzystana informacja o niearbitralnym uporządkowaniu kategorii jednej bądź kilku z analizowanych zmiennych. Wykorzystanie tej informacji pozwala sformułować dodatkowe modele opisujące związki pomiędzy zmiennymi.

Rozważmy na początku sytuację, gdy analizujemy rozkład łączny dwóch zmiennych. O ile zmienne nie są niezależne stochastycznie, należy przyjąć model nasycony, który pełni jedynie funkcję opisową. Z modelem tym nie jest związana żadna hipoteza dotycząca związku pomiędzy zmiennymi, zmienne mogą być zależne, ale nie można nic powiedzieć na temat rodzaju tej zależności. Model nasycony posiada zero stopni swobody. Jak zostanie pokazane w tym rozdziale wykorzystanie informacji o porządkowym charakterze jednej lub dwóch zmiennych, pozwoli na wyszczególnienie pewnych typów zależności, do opisania których potrzeba mniej parametrów aniżeli w modelu nasyconym. W tym sensie zaprezentowane modele dla dwóch zmiennych będą modelami „pośrednimi" pomiędzy niezależnością stochastyczną a modelem nasyconym. Modele te opisują sytuację, w której zmienne są od siebie zależne, natomiast zależność ta ma pewien specyficzny charakter. Związki pomiędzy zmiennymi w modelach dla zmiennych porządkowych formułuje się na ogół za pomocą stosunków szans.

Na początku zostaną przedstawione własności stosunków szans oraz pojęcie niemniejszości stochastycznej. Pojęcia te są pomocne przy analizowaniu związków pomiędzy zmiennymi porządkowymi. Następnie przedstawione zostaną modele formułowane dla dwóch zmiennych, z których jedna bądź obie mierzone są na skali porządkowej. Omówione zostaną hipotezy związane z tymi modelami, kwestie ich parametryzacji

i estymacji. W dalszej części pokazane zostanie, że analogicznie modele formułować można dla większej liczby zmiennych.

## 2.1 Lokalne stosunki szans i ich własności

W rozdziale pierwszym omówione zostały *stosunki szans* opisujące zależności pomiędzy dowolnymi kategoriami zmiennych. W sytuacji, gdy analizowany jest rozkład łączny zmiennej $X$, która przyjmuje $r$ wartości oraz zmiennej $Y$, która przyjmuje $c$ wartości można wyróżnić $\binom{r}{2}\binom{c}{2}$ różnych stosunków szans. Okazuje się jednak, że określony w ten sposób zbiór jest „nadmiarowy" w takim sensie, że na podstawie wybranych stosunków szans jesteśmy w stanie odtworzyć pozostałe. Taki „minimalny" zbiór stosunków szans można określić na wiele sposobów. Zazwyczaj wykorzystuje się na tzw. *lokalne stosunki szans*, które wyróżnia się dla sąsiednich kategorii jednej i drugiej zmiennej. Daje się pokazać, że zestaw lokalnych stosunków szans:

$$\Theta_{ij}^{XY} = \Theta_{(i+1)/i;(j+1)/j}^{X\ \ Y} = \frac{\pi_{ij}^{XY} \cdot \pi_{(i+1)(j+1)}^{XY}}{\pi_{i(j+1)}^{XY} \cdot \pi_{(i+1)j}^{XY}} \tag{2.1}$$

określony dla wszystkich $i$=1, 2,..., $r-1$ oraz $j$=1, 2,..., $c-1$ kategorii, pozwala na odtworzenie wszystkich pozostałych stosunków szans dla zmiennych $X$ i $Y$. Każdy stosunek szans daje się przedstawić jako iloczyn odpowiednich lokalnych stosunków szans. Formalnie:

$$\Theta_{(a+k)/a;(b+m)/b}^{X\ \ Y} = \prod_{i=a}^{a+k-1} \prod_{j=b}^{b+m-1} \Theta_{ij}^{XY} \tag{2.2}$$

dla każdego $a$, $b$, $k$, $m$, takich, że $(a+k) \leq (r-1)$, $(b+m) \leq (c-1)$. Wartości $k$, $m$ wskazują o ile kategorii od siebie znajdują się porównywane wartości zmiennej wierszowej i kolumnowej. Aby lepiej wyjaśnić powyższy zapis formalny pomocny będzie przykład: stosunek szans dla kategorii $x_b$ i $x_d$ zmiennej $X$ oraz $y_f$ i $y_h$ zmiennej $Y$ dla rozkładu z tabeli 2.1 daje się przedstawić jako:

$$\begin{aligned}\Theta_{d/b;h/f}^{X\ \ Y} = \Theta_{c/b;g/f}^{X\ \ Y} \cdot \Theta_{c/b;h/g}^{X\ \ Y} \cdot \Theta_{d/c;g/f}^{X\ \ Y} \cdot \Theta_{d/c;h/g}^{X\ \ Y} = \\ = \frac{\pi_{bf}^{XY} \cdot \pi_{cg}^{XY}}{\pi_{bg}^{XY} \cdot \pi_{cf}^{XY}} \cdot \frac{\pi_{bg}^{XY} \cdot \pi_{ch}^{XY}}{\pi_{bh}^{XY} \cdot \pi_{cg}^{XY}} \cdot \frac{\pi_{cf}^{XY} \cdot \pi_{dg}^{XY}}{\pi_{cg}^{XY} \cdot \pi_{df}^{XY}} \cdot \frac{\pi_{cg}^{XY} \cdot \pi_{dh}^{XY}}{\pi_{ch}^{XY} \cdot \pi_{dg}^{XY}} = \\ = \frac{\pi_{bf}^{XY} \cdot \pi_{dh}^{XY}}{\pi_{bh}^{XY} \cdot \pi_{df}^{XY}}\end{aligned}$$

Zestaw wszystkich lokalnych stosunków szans zdefiniowany w 2.1 opisuje więc w pełni związki pomiędzy kategoriami obydwu zmiennych $X$ i $Y$. Należy zauważyć, że lokalne stosunki szans będą szczególnie użyteczne do opisu związku pomiędzy

Tabela 2.1: Rozkład łączny dwóch zmiennych $X$ i $Y$ (przykład omawiany w teście)

| $X\backslash Y$ | $y_e$ | $y_f$ | $y_g$ | $y_h$ | $y_i$ |
|---|---|---|---|---|---|
| $x_a$ | $\pi_{ae}^{XY}$ | $\pi_{af}^{XY}$ | $\pi_{ag}^{XY}$ | $\pi_{ah}^{XY}$ | $\pi_{ai}^{XY}$ |
| $x_b$ | $\pi_{be}^{XY}$ | $\pi_{bf}^{XY}$ | $\pi_{bg}^{XY}$ | $\pi_{bh}^{XY}$ | $\pi_{bi}^{XY}$ |
| $x_c$ | $\pi_{ce}^{XY}$ | $\pi_{cf}^{XY}$ | $\pi_{cg}^{XY}$ | $\pi_{ch}^{XY}$ | $\pi_{ci}^{XY}$ |
| $x_d$ | $\pi_{de}^{XY}$ | $\pi_{df}^{XY}$ | $\pi_{dg}^{XY}$ | $\pi_{dh}^{XY}$ | $\pi_{di}^{XY}$ |

zmiennymi porządkowymi. W przypadku tych zmiennych kolejność kategorii nie jest ustalona arbitralnie, tak więc rozpatrywanie stosunku szans dla sąsiednich kategorii ma pewien merytoryczny sens. W dalszej części tego rozdziału pokazane zostanie, że lokalne stosunki można wykorzystać do formułowania hipotez dotyczących rozkładu łącznego zmiennych porządkowych.

## 2.2 Niemniejszość stochastyczna

Pojęcie to pozwala porównywać ze sobą rozkład dwóch zmiennych mierzonych na skali porządkowej (lub mocniejszej), bądź porównywać dwa rozkłady warunkowe jednej zmiennej w różnych podzbiorowościach. Dla nas szczególnie użyteczne będzie drugie z wymienionych zastosowanie tego pojęcia. Niemniejszość stochastyczna bazuje na porównaniu rozkładu skumulowanego zmiennej *do k-tej wartości*, tj.

$$P(X \leq x_k) = \sum_{i=1}^{k} \pi_i^X$$

lub rozkładu skumulowanego *od k-tej wartości*:

$$P(X \geq x_k) = \sum_{i=k}^{r} \pi_i^X.$$

Zmienna $X$ ma w podzbiorowości $Y{=}y_a$ rozkład nie mniejszy stochastycznie niż w podzbiorowości $Y{=}y_b$ jeśli rozkład prawdopodobieństwa tej zmiennej spełnia następujący warunek:

$$(X|Y = y_a) \geq_{st} (X|Y = y_b) \Leftrightarrow P(X \leq x_k|Y = y_a) \leq P(X \leq x_k|Y = y_b), \qquad (2.3)$$

bądź równoważnie

$$(X|Y = y_a) \geq_{st} (X|Y = y_b) \Leftrightarrow P(X \geq x_k|Y = y_a) \geq P(X \geq x_k|Y = y_b). \qquad (2.4)$$

dla każdej wartości $x_k$ zmiennej $X$. Oznacza to, że porównując dwie podzbiorowości „wyższe" wartości zmiennej $X$ występują relatywnie częściej w podzbiorowości $Y = y_a$ niż w podbiorowości $Y = y_b$. Tabela 2.2 stanowi ilustrację rozkładów pozostających w relacji nie mniejszości stochastycznej. W górnej części tabeli zamieszczony został rozkład łączny dwóch zmiennych *ocena sytuacji materialnej* i *wykształcenie*, w dolnej części tabeli przedstawiono warunkowe odsetki skumulowane. Porównajmy osoby z wykształceniem wyższym z osobami z wykształceniem średnim. Jak widać, własną sytuację materialną jako „bardzo dobrą" określa 15% osób z wykształceniem wyższym i jedynie 10% osób z wykształceniem średnim. Jako „dobrą" lub „bardzo dobrą" opisuje swoją sytuację 55% osób z wykształceniem wyższym i 25% osób z wykształceniem średnim. Podobnie jest, gdy porównamy odsetki skumulowane wyróżnione dla innych kategorii. Wynika z tego, że rozkład zmiennej „ocena własnej sytuacji materialnej" jest niemniejszy stochastycznie w podzbiorowości osób z wykształceniem wyższym w porównaniu do osób z wykształceniem średnim.

Tabela 2.2: Wykształcenie a ocena własnej sytuacji materialnej — rozkład łączny (w procentach, dane fikcyjne)

| Rozkład łączny (%) | | | |
|---|---|---|---|
| Ocena własnej sytuacji materialnej ($X$) | Wykształcenie ($Y$) | | |
| | 1. Podstawowe | 2. Średnie | 3. Wyższe |
| 1. Bardzo ciężka | 3,75 | 7,50 | 1,75 |
| 2. Dość ciężka | 6,00 | 10,00 | 3,50 |
| 3. Przeciętna | 3,00 | 20,00 | 10,50 |
| 4. Dość dobra | 1,50 | 7,50 | 14,00 |
| 5. Bardzo dobra | 0,75 | 5,00 | 5,25 |
| Skumulowane rozkłady warunkowe $X$ względem $Y$ od danej wartości (%) | | | |
| Ocena własnej sytuacji materialnej ($X$) | Wykształcenie ($Y$) | | |
| | 1. Podstawowe | 2. Średnie | 3. Wyższe |
| 1. Bardzo ciężka | 100 | 100 | 100 |
| 2. Dość ciężka | 75 | 85 | 95 |
| 3. Przeciętna | 35 | 65 | 85 |
| 4. Dość dobra | 15 | 25 | 55 |
| 5. Bardzo dobra | 5 | 10 | 15 |

Bardziej szczegółowa analiza danych z tabeli 2.2 pokazuje, że relacja niemniejszości stochastycznej zachodzi również jeśli porównamy osoby z wykształceniem średnim i podstawowym. Ponieważ jest to własność przechodnia wynika z tego, że *samoocena sytuacji materialnej* musi być nie mniejsza stochastycznie jeśli porównamy osoby z wykształceniem wyższym i podstawowym. Jeśli relacja niemniejszości stochastycznej zachodzi dla kolejnych kategorii zmiennej porządkowej $Y$ możemy wówczas mówić o *zależności regresyjnej* zmiennej $X$ od zmiennej $Y$. Zmienna $X$ jest pozytywnie zależna regresyjnie od zmiennej $Y$, jeśli:

$$(X|Y = y_{j+1}) \geq_{st} (X|Y = y_j) \qquad \text{dla każdej } j \text{ kategorii } y_j. \tag{2.5}$$

Analogicznie zmienna $X$ jest negatywnie zależna regresyjnie od zmiennej $Y$, jeśli:

$$(X|Y = y_j) \geq_{st} (X|Y = y_{j+1}) \qquad \text{dla każdej } j \text{ kategorii } y_j. \tag{2.6}$$

Zależność regresyjna nie jest własnością symetryczną. Oznacza to, że jest możliwe, że zmienna $X$ jest zależna regresyjnie od zmiennej $Y$ (pozytywnie lub negatywnie), a zmienna $Y$ nie jest zależna regresyjnie od zmiennej $X$ (przeanalizowanie rozkładów warunkowych $Y$ względem $X$ w powyższym przykładzie pokazuje asymetryczność zależności regresyjnej). Jak łatwo zauważyć, mówienie o zależności regresyjnej ma sens tylko wtedy gdy obydwie zmienne są mierzone na skali porządkowej lub mocniejszej.

## 2.3 Modele dla dwóch zmiennych

W tej części przedstawione zostaną modele logarytmiczno–liniowe dla dwóch zmiennych, z których jedna bądź dwie są zmiennymi porządkowymi. Modele te zostały sformułowane przez Goodmana (1979*b*) i Habermana (1974*a*) w latach siedemdziesiątych. Tak jak zostało zasygnalizowane na początku tego rozdziału, modele te w pewnym sensie możemy traktować jako „pośrednie" pomiędzy hipotezą niezależności stochastycznej a modelem nasyconym. Zaprezentowane zostaną: model jednakowej interakcji, model wierszowy i dwie wersje modeli wierszowo–kolumnowych. Druga wersja modelu wierszowo–kolumnowego daje możliwość skalowania kategorii zmiennej. Tak jak przy omawianiu modeli dla zmiennych nominalnych zostaną zaprezentowane hipotezy dotyczące rozkładu łącznego związane z poszczególnymi modelami, kwestia ich parametryzacji, estymacji oraz weryfikacji statystycznej.

### 2.3.1 Model jednakowej interakcji (UA)

Model ten dotyczy dwóch zmiennych, z których każda mierzona jest na skali porządkowej lub mocniejszej. Przyjmujemy więc, że:

$$x_1 < x_2 < \ldots < x_j < \ldots < x_r \qquad oraz \qquad y_1 < y_2 < \ldots < y_j < \ldots < y_c.$$

Do opisu związku pomiędzy zmiennymi wykorzystuje się pojęcie lokalnego stosunku szans. Jak pamiętamy, model niezależności stochastycznej zakłada, że wszystkie stosunki szans są równe 1. W modelu nasyconym, gdy na rozkład dwóch zmiennych nie nakładamy żadnych ograniczeń wartości stosunków szans mogą być dowolne. Hipoteza o *jednakowej interakcji* (oznaczana jako UA — uniform association) głosi, że wszystkie lokalne stosunki szans, tj. stosunki szans wyznaczone dla sąsiednich kategorii zmiennej X i zmiennej Y są sobie równe, tj.

$$\Theta_{ij}^{XY} = \delta \qquad \text{dla każdego } i = 1, 2, \ldots, r-1, \qquad \text{oraz } j = 1, 2, \ldots, c-1. \tag{2.7}$$

Lokalny stosunek szans, tj. wyrażenie $\Theta_{ij}^{XY}$ zostało zdefiniowane w formule 2.1. Hipotezę o jednakowej interakcji ilustruje tabela 2.3 (wielkości w tej tabeli zostaną omówione poniżej).

Tabela 2.3: Rozkład łączny zmiennych $X$ i $Y$ ilustrujący model *jednakowej interakcji*

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $x_1$ | $\gamma$ | $\gamma \cdot \gamma_2^Y$ | $\gamma \cdot \gamma_3^Y$ | $\gamma \cdot \gamma_4^Y$ |
| $x_2$ | $\gamma \cdot \gamma_2^X$ | $\gamma \cdot \gamma_2^X \cdot \gamma_2^Y \cdot \delta$ | $\gamma \cdot \gamma_2^X \cdot \gamma_3^Y \cdot \delta^2$ | $\gamma \cdot \gamma_2^X \cdot \gamma_4^Y \cdot \delta^3$ |
| $x_3$ | $\gamma \cdot \gamma_3^X$ | $\gamma \cdot \gamma_3^X \cdot \gamma_2^Y \cdot \delta^2$ | $\gamma \cdot \gamma_3^X \cdot \gamma_3^Y \cdot \delta^4$ | $\gamma \cdot \gamma_3^X \cdot \gamma_4^Y \cdot \delta^6$ |
| $x_4$ | $\gamma \cdot \gamma_4^X$ | $\gamma \cdot \gamma_4^X \cdot \gamma_2^Y \cdot \delta^3$ | $\gamma \cdot \gamma_4^X \cdot \gamma_3^Y \cdot \delta^6$ | $\gamma \cdot \gamma_4^X \cdot \gamma_4^Y \cdot \delta^9$ |

Hipoteza ta opisuje pewien szczególny typ zależności między zmiennymi. Istotne jest, że jeśli hipoteza ta jest spełniona, zależność — opisywana lokalnym stosunkiem szans — jest taka sama, dla każdej tabeli o wymiarach 2 x 2 złożonej z sąsiednich kategorii zmiennej $X$ i sąsiednich kategorii zmiennej $Y$. Siłę tak zdefiniowanego związku definiuje wielkość lokalnego stosunku szans czyli wielkość $\delta$. Jeśli $\delta > 1$ to z własności (2.2) wiadomo, że wszystkie stosunki szans wyznaczone dla kategorii $x_a$ oraz $x_b$ zmiennej $X$ oraz kategorii $y_c$, $y_d$ zmiennej $Y$ takich, że $a < b$ oraz $c < d$, będą większe od 1, tj. $\Theta_{a/b;c/d}^{X\ Y} > 1$, bez względu na to ile kategorii dzieli wartości $x_a$, $x_b$ oraz $y_c$, $y_d$ (przy czym dla kategorii bardziej oddalonych od siebie odpowiedni stosunek szans jest większy).

Sytuacja gdy $\delta > 1$ wskazuje więc na pozytywną zależność pomiędzy zmiennymi porządkowymi. Im większa jest wartość lokalnego stosunku szans, tym ta zależność pomiędzy zmiennymi silniejsza jest silniejsza. Jeśli $\delta < 1$ wskazuje to na zależność negatywną [1]. W sytuacji niezależności stochastycznej wielkość $\delta$ byłaby równa 1.

Istnieje związek pomiędzy hipotezą o jednakowej interakcji a stochastyczną niemniejszością rozkładów warunkowych jednej zmiennej względem drugiej zmiennej. Jeśli rozkład jest zgodny z modelem jednakowej interakcji, to zmienna $X$ jest zależna regresyjnie od zmiennej $Y$ jak również zmienna $Y$ jest zależna regresyjnie od zmiennej $X$. Kierunek tej zależności — pozytywny lub negatywny — zależy wielkości lokalnego stosunku szans. Jeśli jest on większy od 1 mamy do czynienia z pozytywną zależnością regresyjną, czyli porównując dwie kategorie $x_a$ oraz $x_b$, takie , że $a < b$, rozkład warunkowy $Y$ jest nie mniejszy stochastycznie w podzbiorowości $x_b$. Można to łatwo zauważyć porównując dowolne dwa wiersze w tabeli 2.3, będącej ilustracją rozkładu zgodnego z hipotezą o jednakowej interakcji. Do podobnych wniosków dojdziemy porównując dowolne dwie kolumny tj. dwa rozkłady warunkowe $X$.

Należy podkreślić, że związek pomiędzy hipotezą o jednakowej interakcji z zależnością regresyjną zachodzi w jedną stronę. Nawet jeśli rozkład łączny dwóch zmiennych spełniać będzie dwa warunki tj. zależność regresyjną $X$ względem $Y$ oraz zależność regresyjną $Y$ względem $X$, to nie oznacza to jeszcze, że wszystkie lokalne stosunki szans są równe, tak więc nie musi być spełniona hipoteza o jednakowej interakcji.

W modelu jednakowej interakcji stosunek szans dla dowolnych kategorii można przedstawić — zgodnie z 2.2 — jako iloczyn odpowiednich lokalnych stosunków szans $\Theta_{ij}^{XY}$. Ponieważ wartości lokalnych stosunków szans są sobie równe można pokazać, że stosunek szans obliczony dla $i$–tej względem pierwszej kategorii zmiennej $X$ a także $j$–tej względem pierwszej kategorii zmiennej Y wynosi:

$$\Theta_{i/1;j/1}^{X\ Y} = \delta^{(i-1)(j-1)}. \tag{2.8}$$

Zamieszczona powyżej tabela 2.3, stanowiąca ilustrację modelu jednakowej interakcji jest analogiczna do tabeli 1.20 zgodnej z parametryzacją względem kategorii odniesienia. Jak widać wszystkie lokalne stosunki szans są sobie równe. Ponieważ lokalny stosunek szans opisuje zależność pomiędzy dwiema zmiennymi, istnieje ścisła relacja pomiędzy parametrem interakcji $\gamma_{ij}^{XZ}$ z tabeli 1.20 a wyrażeniem $\delta^{(i-1)(j-1)}$.

[1] Należy zauważyć, że jest to zależność innego rodzaju niż opisują np. mierniki korelacji rangowej. Na przykład używany często współczynnik $tau - b$ Kendalla przyjmuje wartość maksymalną, gdy jedna zmienna jest rosnącą funkcją drugiej zmiennej. W tym przypadku rozkład nie byłby zgodny z hipotezą o jednakowej interakcji.

Tabela 2.4: Tablica ilustrująca model ilustrujący model *jednakowej interakcji* — parametry opisujące związek między zmiennymi

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\delta$ | $\delta^2$ | $\delta^3$ |
| $x_3$ | 1 | $\delta^2$ | $\delta^4$ | $\delta^6$ |
| $x_4$ | 1 | $\delta^3$ | $\delta^6$ | $\delta^9$ |

Tabela 2.4 jest uproszczoną wersją poprzedniej tabeli, tj. zawiera wyłącznie parametry interakcji i pomija parametry $\gamma$, $\gamma_i^X$, $\gamma_j^Y$, które nie opisują związku pomiędzy zmiennymi a jedynie rozkłady poszczególnych zmiennych. Ten sposób prezentacji wykorzystany zostanie przy wprowadzaniu kolejnych modeli w dalszej części tej racy. W obydwu tabelach 2.3, 2.4 podobnie jak w tabeli 1.20 kategoriami odniesienia są pierwsze kategorie obydwu zmiennych. Dlatego też wyrażenie $\delta^{(i-1)(j-1)}$ jest stosunkiem szans wyznaczonym względem pierwszych kategorii obydwu zmiennych. Oczywiście, możliwe jest przyjęcie innych kategorii odniesienia. Ogólnie, jeśli kategoriami odniesienia dla obydwu zmiennych są odpowiednio wartości $x_a$ oraz $y_b$ to model zgodny z tą parametryzacją przyjmuje postać:

$$\pi_{ij}^{XY} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \delta^{(i-a)(j-b)}. \tag{2.9}$$

Przy czym parametry $\gamma_a^X$, $\gamma_b^Y$, są równe 1 zgodnie z 1.53. Model jednakowej interakcji w parametryzacji odchyleń multiplikatywnych różni się nieznacznie:

$$\pi_{ij}^{XY} = \tau \cdot \tau_i^X \cdot \tau_j^Y \cdot \delta^{(i-\bar{i})(j-\bar{j})}. \tag{2.10}$$

Stałe $\bar{i}$ oraz $\bar{j}$ dobrane są w taki sposób, że $\sum(i-\bar{i}) = 0$ oraz $\sum(j-\bar{j}) = 0$. Wynika to z warunków 1.36 nakładanych na parametry interakcji przy tym rodzaju parametryzacji, tj. iloczyn parametrów interakcji dla poszczególnych wierszy oraz iloczyn parametrów dla kolumn jest równy 1, tj.

$$\prod_{i=1}^{r} \delta^{(i-\bar{i})(j-\bar{j})} = 1 \qquad \text{oraz} \qquad \prod_{j=1}^{c} \delta^{(i-\bar{i})(j-\bar{j})} = 1. \tag{2.11}$$

Porównując model jednakowej interakcji do modelu niezależności stochastycznej widzimy, że pierwszy z nich wykorzystuje tylko jeden parametr więcej niż drugi. Parametr $\delta$ opisuje związek miedzy zmiennymi i można go interpretować jako lokalny stosunek szans. Ta interpretacja pozostaje taka sama dla obydwu parametryzacji.

Model jednakowej interakcji można przedstawić również w formie addytywnej:

$$\log \pi_{ij}^{XY} = \mu + \lambda_i^X + \lambda_j^Y + a(i - \bar{i})(j - \bar{j}). \tag{2.12}$$

gdzie – jak daje się pokazać – parametr $a$ jest równy logarytmowi parametru $\delta$ z wersji multiplikatywnej, czyli logarytmowi lokalnego stosunku szans. Aby móc jednoznacznie określić metodą największej wiarygodności rozkład oczekiwany zgodny z hipotezą o jednakowej interakcji – oprócz informacji na temat rozkładów brzegowych obydwu zmiennych – potrzebna jest informacja na temat współwystępowania ze sobą różnych kategorii obydwu zmiennych. Innymi słowy szacowane prawdopodobieństwa rozkładu oczekiwanego $\widehat{\pi}_{ij}$ muszą spełniać następujące warunki:

$$\widehat{\pi}_i^X = p_i^X, \tag{2.13}$$

$$\widehat{\pi}_j^Y = p_j^Y, \tag{2.14}$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c} i \cdot j \cdot \widehat{\pi}_{ij}^{XY} = \sum_{i=1}^{r}\sum_{j=1}^{c} i \cdot j \cdot p_{ij}^{XY}, \tag{2.15}$$

gdzie $p_{ij}$ to częstości z próby. Wyrażenie 2.15 informuje nas z jakimi kategoriami jednej zmiennej współwystępują kategorie drugiej zmiennej. Jeśli jest dodatnie, może to wskazywać na zależność pozytywną. Hipoteza dotycząca jednakowej interakcji jest hipotezą nie–elementarną, nie istnieje więc formuła pozwalająca wyznaczyć rozkład oczekiwany zgodny z powyższym modelem. Konieczne jest zastosowanie metod iteracyjnych (Goodman 1979*b*, Agresti 1984).

Jak zostało powiedziane wcześniej model jednakowej interakcji posiada jedynie o jeden parametr więcej aniżeli model niezależności stochastycznej - parametr $\delta$. Liczba stopni swobody tego modelu wynosi więc: $df = (r-1)(c-1) - 1 = rc - r - c$.

Liczbę stopni swobody można zdefiniować również odwołując się do liczby niezależnych od siebie lokalnych stosunków szans. Dla tabeli o wymiarach $r$ x $c$ można wyznaczyć $(r-1)(c-1)$ takich wielkości. Zgodnie z omawianym modelem, wszystkie one są sobie równe, tj.

$$\Theta_{11}^{XY} = \Theta_{12}^{XY} = \ldots = \Theta_{(r-1)(c-1)}^{XY},$$

a więc liczba odpowiednich niezależnych od siebie warunków wynosi $df = (r-1)(c-1) - 1$.

W tabeli 2.5 przedstawiony został rozkład łączny dwóch zmiennych: ocena własnej sytuacji materialnej i własnego gospodarstwa domowego $(X)$ oraz opinie dotyczące tego, czy rząd powinien zmniejszyć różnice w dochodach$(Y)$[2]. Dane pochodzą

[2]Pytanie o sytuację materialną brzmiało: *Interesuje nas również, jak obecnie ludzie radzą sobie finansowo. Biorąc pod uwagę swoją sytuację i sytuację finansową Pana(-i) rodziny, proszę powiedzieć,*

Tabela 2.5: Zadowolenie z własnej sytuacji materialnej a opinie dotyczące zmniejszenia różnic w dochodach przez rząd[a]

| Rząd powinien zmniejszyć różnice w dochodach ($X$) | Zadowolenie z własnej sytuacji finansowej ($Y$) | | |
|---|---|---|---|
| | 1. Zadowolony(a) | 3. Mniej więcej zadowolony(a) | 3. Niezadowolony(a) 3 |
| 1. Zdecydowanie się zgadzam | 53,2 | 160,8 | 366,3 |
| 2. Zgadzam się | 53,8 | 195,1 | 235,2 |
| 3. Ani się zgadzam, ani nie zgadzam | 14,2 | 43,6 | 36,6 |
| 4. Nie zgadzam się | 15,1 | 24,3 | 22,4 |
| 5. Zdecydowanie się nie zgadzam | 3,9 | 10,6 | 2,0 |

[a]Źródło: Polski Generalny Sondaż Społeczny, 2005. Dane przeważone.

Tabela 2.6: Wyniki weryfikacji dla danych z tabeli 2.5

| Model | df | $\chi^2$ | $G^2$ | $\Delta$ |
|---|---|---|---|---|
| $[X][Y]$ | 8 | 60,2 (p<0,0001) | 59,7 (p<0,0001) | 9,0 |
| UA | 7 | 13,5 (p=0,0604) | 13,5 (p=0,0604) | 4,2 |
| R | 4 | 12,8 (p=0,0124) | 12,7 (p=0,0130) | 3,8 |
| C | 6 | 7,1 (p=0,3097) | 7,2 (p=0,3035) | 2,7 |
| RC1 | 3 | 6,7 (p=0,0831) | 6,7 (p=0,0823) | 2,5 |
| RC2 | 3 | 4,7 (p=0,1931) | 4,5 (p=0,2089) | 1,4 |

z Polskiego Generalnego Sondażu Społecznego, z 2005 roku (Cichomski i inni, 2009). W tabeli 2.6 zostały przedstawione wyniki dopasowania do danych dla kilku modeli.

*czy jest Pan(-i), zadowolony, mniej więcej zadowolony, czy też niezadowolony ze swojej obecnej sytuacji finansowej* . W drugim pytaniu respondenci mieli się ustosunkować do opinii: *Do zadań rządu powinno należeć zmniejszenie różnic pomiędzy wysokimi i niskimi dochodami*. Respondenci mieli do wyboru odpowiedzi podane w tabeli 2.5. Osoby, które nie udzieliły odpowiedzi na przynajmniej jedno z pytań zostały wyłączone z analizy (stanowiły one 2,4% wszystkich respondentów). Zastosowano złożony schemat doboru próby, jednak — tak jak zostało zasygnalizowane w poprzednim rozdziale — weryfikacja hipotez w tej i w kolejnych analizach została przeprowadzona przy założeniu, że mamy do czynienia z doborem prostym. Dane zostały przeważone, dlatego liczebności w tabeli nie są liczbami całkowitymi. Więcej informacji o badaniu można znaleźć w Aneksie i na stronie internetowej *pgss.iss.uw.edu.pl*.

Część z nich omówiona zostanie w dalszej części tego rozdziału. Modelem wyjściowym jest model niezależności stochastycznej. Jak widać, model ten nie jest realistycznym opisem danych. Okazuje się jednak, że model jednakowej interakcji (UA) — choć posiada tylko jeden dodatkowy parametr — jest akceptowalny na poziomie istotności $\alpha = 0,05$. Jego liczba stopni swobody wynosi 7, a statystyki dopasowania wynoszą odpowiednio $\chi^2 = 13,52$ $(p = 0,06)$, $G^2 = 13,50$ $(p = 0,06)$.

Tabela 2.7: Rozkład oczekiwany zgodny z modelem jednakowej interakcji dla danych z tabeli 2.5

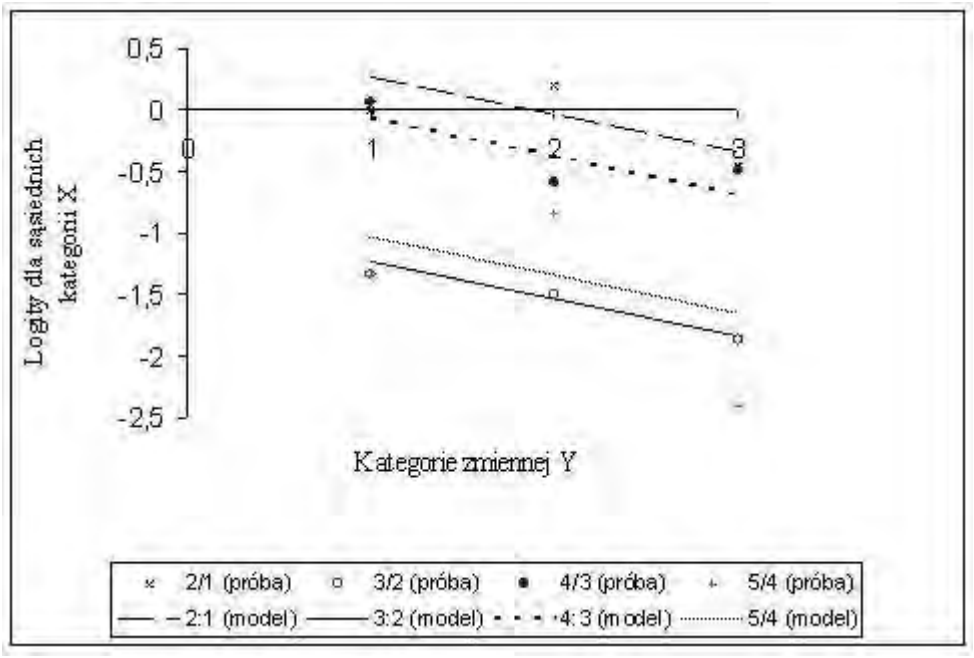
| Rząd powinien zmniejszyć różnice w dochodach ($X$) | Zadowolenie z własnej sytuacji finansowej ($Y$) | | |
|---|---|---|---|
| | 1. Zadowolony(a) | 3. Mniej więcej zadowolony(a) | 3. Niezadowolony(a) 3 |
| 1. Zdecydowanie się zgadzam | 44,2 | 185,0 | 351,1 |
| 2. Zgadzam się | 57,7 | 177,9 | 248,4 |
| 3. Ani się zgadzam, ani nie zgadzam | 16,9 | 38,2 | 39,3 |
| 4. Nie zgadzam się | 15,7 | 26,2 | 19,8 |
| 5. Zdecydowanie się nie zgadzam | 5,7 | 7,0 | 3,9 |

W tabeli 2.7 zostały podane liczebności oczekiwane zgodne z omawianym modelem. Jak łatwo sprawdzić wszystkie lokalne stosunki są sobie równe:

$$\delta = \frac{44,2 \cdot 177,9}{185,0 \cdot 57,7} = \frac{185,0 \cdot 248,4}{351,1 \cdot 177,9} = \ldots = 0,74.$$

Zinterpretujmy powyższy parametr w odniesieniu do stosunku szans $\Theta_{21}^{XY}$. Zgodnie z modelem jednakowej interakcji proporcja liczby osób, które *zgadzają się* co do tego, że rząd powinien zmniejszyć różnice w dochodach $(X = x_2)$ do liczby osób, które są *neutralne* wobec takiej opinii$(X = x_3)$ jest 1,35 (1/0,74) razy mniejsza wśród respondentów, którzy są *zadowoleni* z sytuacji finansowej $(Y = y_1)$ aniżeli wśród osób *średnio zadowolonych* $(Y = y_2)$. Podobny wniosek można sformułować dla dowolnej pary „sąsiednich" kategorii zmiennej $X$ i zmiennej $Y$. Jest również możliwe określenie zależności dla kategorii „odległych" od siebie, na przykład $\Theta_{5/1;3/1}^{X\ Y} = \delta^8$, co oznacza, że zgodnie z modelem proporcja liczby osób, które *zdecydowanie zgadzają się* co do tego, że rząd powinien zmniejszyć różnice w dochodach $(X = x_1)$ do liczby osób, które *zdecydowanie się nie zgadzają* z taką opinią $(X = x_5)$ jest ponad jedenaście razy

Rysunek 2.1: Warunkowe logity dla sąsiednich kategorii zmiennej $X$ względem $Y$ (model jednakowej interakcji)




mniejsza ($1/0,74$[8]), wśród respondentów, którzy są *zadowoleni* z sytuacji finansowej ($Y = y_1$) aniżeli wśród osób *niezadowolonych* ($Y = y_3$).

Warto kilka słów poświecić graficznej interpretacji modelu jednakowej interakcji. Często w literaturze wskazuje się, że model ten opisuje pewnego rodzaju liniową zależność[3]. Rysunek 2.1 pokazuje, na czym polega liniowość tej zależności w odniesieniu do analizowanego powyżej przykładu empirycznego[4]. Na osi pionowej naniesione zostały warunkowe logarytmy szans[5], czyli tzw. *logity* wyznaczone dla sąsiednich kategorii zmiennej $X$, względem poszczególnych kategorii zmiennej $Y$ tj. $log\left(\Omega^{X\ (Y)}_{2/1(j)}\right)$, $log\left(\Omega^{X\ (Y)}_{3/2(j)}\right)$, $log\left(\Omega^{X\ (Y)}_{4/3(j)}\right)$, $log\left(\Omega^{X\ (Y)}_{5/4(j)}\right)$. Wielkości te wskazują informują jak często była wybierana kategoria $x_{i+1}$ względem kategorii $x_i$. Na przykład, jeśli $log\left(\Omega^{X\ (Y)}_{2/1(1)}\right) = 0$ oznacza to, że osoby zadowolone z własnej sytuacji finansowej, w kwestii zmniejszania różnic w dochodach przez rząd równie często wskazują na odpowiedź *zdecydowanie się zgadzam* jak odpowiedź *zgadzam się*. Wielkość większa od 0 wskazuje, że relatywnie częściej wybierana była odpowiedź *zgadzam się*, natomiast wielkość mniejsza od 0 — odpowiedź *zdecydowanie się zgadzam*.

[3]W literaturze anglosaskiej określa się go często jako *linear–by–linear model* (Agresti 1984, Haberman 1974*a*).

[4]Podobną ilustrację w odniesieniu do tablic ruchliwości zamieszcza Hout (1983).

[5]Pojęcie szansy zostało omówione w rozdziale 1 przy okazji modelu $[XY][XZ][YZ]$.

Na rysunku naniesione zostały zarówno logity z próby, jak również logity zgodne z założonym modelem. Widoczna jest następująca tendencja: wielkości logitów maleją, gdy przechodzimy od osób zadowolonych do osób niezadowolonych, co wskazuje na następujący trend: im gorsza jest sytuacja materialna, tym częściej respondenci postulują zmniejszenie różnic w dochodach. Zauważmy, że logity opisujące jednakową interakcję, układają się na linii prostej a jej kąt nachylenia określa wielkość:

$$log\left(\Omega^{X\ \ (Y)}_{i+1/i(j+1)}\right) - log\left(\Omega^{X\ \ (Y)}_{i+1/i(j)}\right) = log\left(\frac{\Omega^{X\ \ (Y)}_{i+1/i(j+1)}}{\Omega^{X\ \ (Y)}_{i+1/i(j)}}\right) = log(\delta) = a,$$

czyli jest równy parametrowi interakcji z addytywnej postaci modelu 2.12. Co więcej, linie dla poszczególnych logitów są równoległe do siebie. To, na jakiej wysokości jest linia zależy od wielkości wyrażenia

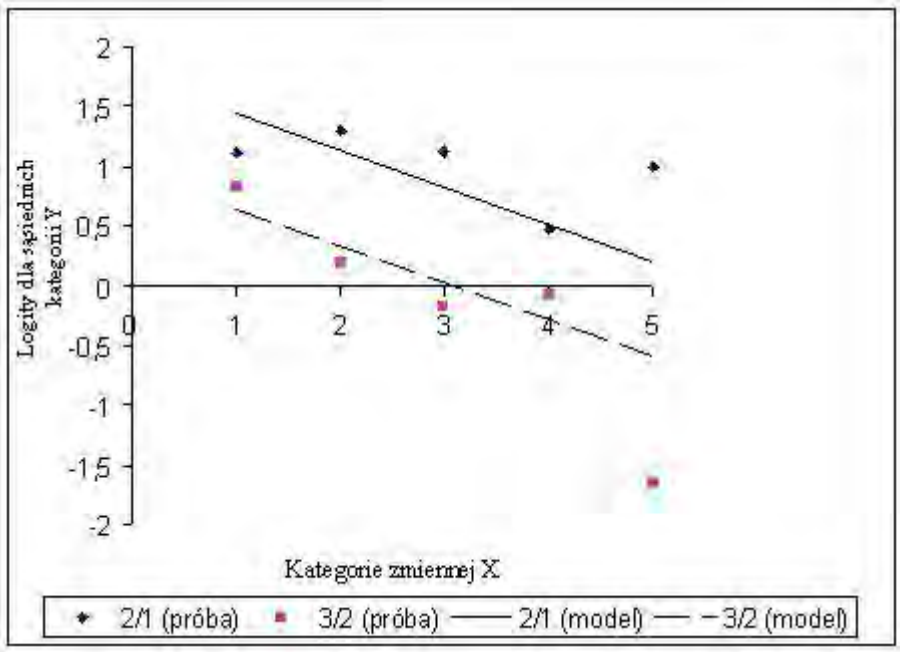
$$\lambda^X_{i+1} - \lambda^X_i.$$

W przypadku parametryzacji odchyleń multiplikatywnych wielkość ta oznacza logarytm cząstkowej szansy — tj. logarytm średniej geometrycznej warunkowych szans — że zmienna $X$ przyjmie raczej wartość $x_{i+1}$ aniżeli wartość $x_i$. Daje się również zauważyć, że logity empiryczne leżą stosunkowo blisko tych linii, co stanowi graficzne potwierdzenie dobrego dopasowania modelu do danych. Stosunkowo najgorzej dopasowane są logity $log\left(\Omega^{X\ (Y)}_{5/4(j)}\right)$, ale warto zauważyć, że liczebności dla piątej kategorii zmiennej $X$ liczebności są małe, stąd można się spodziewać, że zależności ustalone na podstawie próby mogą być najbardziej przypadkowe. Rysunek 2.2 przedstawia analogiczne logity dla sąsiednich kategorii zmiennej $Y$ względem kategorii zmiennej $X$. Wzór zależności jest podobny: logity zgodne z modelem są liniowe względem kolejnych kategorii zmiennej $X$, a linie dla logitów $log\left(\Omega^{Y\ (X)}_{2/1(i)}\right)$ oraz $log\left(\Omega^{Y\ (X)}_{3/2(i)}\right)$ są równoległe do siebie. Również na tym rysunku widać, że najgorzej dopasowane są logity wyznaczone dla podzbiorowości określonej przez piątą wartość zmiennej $X$, co może wynikać z tego, że kategoria ta była wybierana przez respondentów sporadycznie.

### 2.3.2 Model efektu wierszowego (R)

Model ten uwzględnia informacje o niearbitralnym uporządkowaniu kategorii jednej zmiennej. Przyjmijmy, że zmienną tą będzie zmienna *kolumnowa* $Y$. Druga zmienna $X$ — zmienna *wierszowa*— może być mierzona na skali nominalnej. Oczywiście model ten może opisywać również związek pomiędzy dwiema zmiennymi porządkowymi. W takiej sytuacji nie wykorzystujemy informacji o uporządkowaniu kategorii zmiennej wierszowej ($X$), jednak może się okazać, że model ten będzie lepiej opisywał związek między zmiennymi niż hipoteza o jednakowej interakcji.

Rysunek 2.2: Warunkowe logity dla sąsiednich kategorii zmiennej $Y$ względem $X$ (model jednakowej interakcji)




W modelu efektu wierszowego (oznaczonym jako R — *row effect model*) — podobnie w modelu jednakowej interakcji — związek pomiędzy dwiema zmiennymi jest opisywany za pomocą stosunku szans. Hipoteza związana z tym modelem głosi, że stosunki szans dla dwóch dowolnych kategorii zmiennej wierszowej $X$ tj. $x_a$, $x_b$, oraz sąsiadujących ze sobą kategorii zmiennej porządkowej $Y$, tj. $y_j$, $y_{j+1}$ zależą tylko od tego, które wartości zmiennej wierszowej są porównywane, natomiast są takie same dla kolejnych kolumn. Formalnie:

$$\Theta^{X}_{b/a;(j+1)/j} = \delta^{X}_{b/a}. \tag{2.16}$$

Prawidłowość ta zachodzi dla każdej pary kategorii $x_a$, $x_b$ zmiennej X oraz dla każdego $j = 1, 2, \ldots, c-1$. Wielkość stosunku szans informuje nas na ile rozkłady oczekiwane zmiennej porządkowej $Y$ dla dwóch kategorii zmiennej $X$ są podobne lub różnią się od siebie. Jeśli stosunki szans są równe jedności wskazuje to, że warunkowe rozkłady prawdopodobieństwa zmiennej $Y$ wyróżnione ze względu na kategorie $x_a$, $x_b$ zmiennej $X$ są takie same.

Ilustracją modelu wierszowego jest tablica 2.8 (interpretacja wielkości w niej przedstawionych zostanie przedstawiona w dalszej części). Porównując tę tabelę z tablicą 2.4, jak również formuły 2.7 oraz 2.16, można zauważyć, że model jednakowej interakcji jest prostszy w stosunku do modelu efektu wierszowego tj. pierwszy z nich jest szczególnym przypadkiem drugiego. Łatwo daje się pokazać związek pomiędzy

obydwiema hipotezami. Hipoteza o efekcie wierszowym głosi, że jeśli z tabeli rozkładu łącznego zmiennych $X$ i $Y$ o wymiarach $r$ x $c$ wyodrębnimy tabelę o wymiarach 2 x $c$, (dla dowolnych dwóch wierszy), to rozkład w tej tabeli będzie zgodny z hipotezą o jednakowej interakcji, bo dla takiej „zredukowanej tabeli" wszystkie lokalne stosunki szans są sobie równe.

Tabela 2.8: Ilustracja modelu *efektu wierszowego* - parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\phi_2$ | $\phi_2^2$ | $\phi_2^3$ |
| $x_3$ | 1 | $\phi_3$ | $\phi_3^2$ | $\phi_3^3$ |
| $x_4$ | 1 | $\phi_4$ | $\phi_4^2$ | $\phi_4^3$ |

Podobnie jak w przypadku modelu jednakowej interakcji, również w przypadku modelu efektu wierszowego można pokazać związek z niemniejszością stochastyczną wybranych rozkładów. Dwa dowolne warunkowe rozkłady zmiennej porządkowej $Y$ w zbiorowościach wyróżnionych, ze względu na kategorie $x_a$, $x_b$, zmiennej wierszowej $X$ pozostają w relacji niemniejszości stochastycznej. To, który rozkład jest nie mniejszy stochastycznie zależy od wielkości lokalnego stosunku szans $\delta_{b/a}$. Jeśli jest on większy od 1, to rozkład dla podzbiorowości $x_b$ jest nie mniejszy stochastycznie w porównaniu do rozkładu w podzbiorowości $x_a$. Zależność ta nie zachodzi jednak w obydwie strony: z tego, że rozkłady warunkowe dla dowolnej pary wierszy pozostają w relacji niemniejszości stochastycznej nie wynika, że rozkład łączny jest zgodny z modelem wierszowym.

Zakładając, że model o efekcie wierszowym jest spełniony, można wyróżnić $\binom{r}{2}$ różnych stosunków szans dla sąsiednich kategorii zmiennej porządkowej $Y$ i dwóch dowolnych kategorii zmiennej $X$, tj. związek pomiędzy zmiennymi opisywałyby $\binom{r}{2}$ parametry $\delta_{b/a}$. Z własności 2.2 wynika jednak, że do opisu zależności pomiędzy zmiennymi nie jest konieczne rozważanie ich wszystkich. Można skupić się na lokalnych stosunkach szans. Hipoteza o efekcie wierszowym głosi, że:

$$\Theta_{ij}^{XY} = \delta_i^X. \tag{2.17}$$

Powyższą hipotezę, możemy również wyrazić w odniesieniu do zestawu $(r-1)$ stosunków szans wyodrębnionych względem kategorii odniesienia $x_1$ zmiennej $X$, tj.

$$\Theta_{i/1;(j+1)/j}^{X\ \ Y} = \delta_{i/1}^X. \tag{2.18}$$

dla każdego $i = 1, 2, \ldots, r-1$ takiego, że $i \neq 1$ oraz $j = 1, 2, \ldots, c-1$. Powyższe formuły są równoważne. Na podstawie zestawu z formuły 2.17 bądź formuły 2.18 jesteśmy w stanie odtworzyć wszystkie stosunki szans, dla dowolnych kategorii zmiennej $X$ i zmiennej $Y$. Na przykład dla dowolnych kategorii, $x_e$, $x_f$ zmiennej $X$ i dowolnego $j$ zachodzi:

$$\Theta^{X\ Y}_{f/e;(j+1)/j} = \frac{\Theta^{X\ Y}_{f/1;(j+1)/j}}{\Theta^{X\ Y}_{e/1;(j+1)/j}} = \frac{\delta^X_{f/1}}{\delta^X_{e/1}} = \delta_{f/e}. \tag{2.19}$$

Tak więc, do porównania rozkładów warunkowych zmiennej $Y$ w kategoriach wyróżnionych ze względu na zmienną wierszową $Y$ wystarczy w przypadku hipotezy o efekcie wierszowym $r-1$ parametrów.

Zauważmy, że tabela 2.8 ilustrująca model efektu wierszowego — podobnie jak 2.4 dotycząca jednakowej interakcji — jest zgodna z parametryzacją względem kategorii odniesienia, przy czym kategoriami odniesienia dla obydwu zmiennych są pierwsze kategorie zmiennych $X$ oraz $Y$. Parametr $\phi_i$ można interpretować jako stosunek szans $\Theta^{X\ Y}_{i/1;(j+1)/j}$. Jest on indeksowany przez wartość zmiennej wierszowej, tak więc do opisu zależności pomiędzy zmiennymi potrzebnych jest więcej parametrów niż w modelu jednakowej interakcji a mianowicie $r-1$. Korzystając z 2.2 wiadomo, że jeśli rozkład jest zgodny z modelem wierszowym to każdy stosunek szans obliczony dla kategorii $j$-tej względem pierwszej kategorii zmiennej $Y$ oraz kategorii $i$-tej względem pierwszej kategorii zmiennej $X$, daje się przedstawić jako iloczyn odpowiednich stosunków szans w danym wierszu:

$$\Theta^{X\ Y}_{i/1;j/1} = \phi_i^{j-1} \qquad \text{dla każdego } i = 1, 2, \ldots, r-1 \text{ oraz } j = 1, 2, \ldots, c-1. \tag{2.20}$$

Oczywiście, można wybrać dowolne kategorie odniesienia $x_c$ oraz $y_d$ obydwu zmiennych. Ogólnie, model efektu wierszowego można przestawić jako:

$$\pi^{XY}_{ij} = \gamma \cdot \gamma^X_i \cdot \gamma^Y_j \cdot \phi_i^{(j-d)}. \tag{2.21}$$

Przy czym parametry $\gamma^X_c$, $\gamma^Y_d$, $\phi_c$ są równe 1 zgodnie z 1.53. Różnica $(j-d)$ informuje, ile kategorii dzieli kategorię $y_j$ od kategorii odniesienia $y_d$. Dla parametryzacji odchyleń multiplikatywnych powyższy model przyjmuje postać:

$$\pi^{XY}_{ij} = \tau \cdot \tau^X_i \cdot \tau^Y_j \cdot \varphi_i^{(j-\bar{j})}. \tag{2.22}$$

Przy czym na parametry $\tau^X_i$, $\tau^Y_j$ nakładamy warunki 1.36, natomiast $\bar{j}$ jest taką stałą, że $\sum(j-\bar{j}) = 0$ a iloczyn wszystkich parametrów $\delta_i$ jest równy 1, tj:

$$\prod_{i=1}^{r} \varphi_i = 1. \tag{2.23}$$

Model efektu wierszowego w formie addytywnej, wygląda następująco:

$$\log \pi_{ij}^{XY} = \mu + \lambda_i^X + \lambda_j^Y + b_i(j - \bar{j}). \tag{2.24}$$

Oszacowanie metodą największej wiarygodności rozkładu oczekiwanego zgodnego z hipotezą o efekcie wierszowym, wymaga oprócz informacji na temat rozkładów brzegowych obydwu zmiennych (2.13, 2.14) również informacji na temat rozkładów warunkowych zmiennej $Y$ względem zmiennej $X$. Mówiąc dokładniej musi zachodzić:

$$\sum_{j=1}^{c} j \cdot \widehat{\pi}_{ij}^{XY} = \sum_{j=1}^{c} j \cdot p_{ij}^{XY}, \tag{2.25}$$

dla każdego $i$, gdzie $p_{ij}$ to częstości z próby. Wyrażenie to informuje nas, czy kategorie „wyższe" zmiennej porządkowej $Y$ pojawiają się częściej czy też rzadziej w stosunku do kategorii „niższych". Podobnie jak w modelu o jednakowej interakcji nie istnieje formuła pozwalająca na skonstruowanie rozkładu oczekiwanego i konieczne jest zastosowanie metod iteracyjnych.

Jak widać, model efektu wierszowego posiada więcej parametrów aniżeli model jednakowej interakcji. W porównaniu do modelu niezależności stochastycznej posiada on dodatkowo $(r-1)$ parametrów $\phi^i$, dlatego liczba stopni swobody tego modelu wynosi:

$$df = (r-1)(c-1) - (r-1) = (r-1)(c-2).$$

Liczbę stopni swobody można również wyznaczyć odwołując się do założeń dotyczących lokalnych stosunków szans. Liczba sąsiadujących ze sobą par wierszy wynosi $(r-1)$, dla każdej pary wierszy można wyznaczyć $(c-1)$ lokalnych stosunków szans. Zakładamy, że dla każdej pary wierszy są one równe, co przekłada się na $(c-2)$ warunków. Łącznie liczba założeń wynosi więc $(r-1)(c-2)$. Na ich podstawie możemy wyznaczyć wartość każdego nie–lokalnego stosunku szans.

Analogicznie daje się sformułować hipotezę o efekcie kolumnowym (oznaczaną jako C — column effect model): zmienna wierszowa jest zmienną porządkową i porównujemy jej rozkłady dla wybranych kategorii zmiennej wierszowej. Liczba stopni swobody tego modelu wynosi $df = (r-2)(c-1)$.

Hipotezę efektu wierszowego wyraża się również za pomocą nieco odmiennej parametryzacji ułatwiającej jej porównanie z modelem jednakowej interakcji. Formułę 2.17 definiującą lokalny stosunek szans można alternatywnie przedstawić jako iloczyn parametrów dwóch typów.

$$\Theta_{ij}^{XY} = \delta \cdot \delta_i \tag{2.26}$$

dla każdego $i = 1, 2, \ldots, r-1$ oraz $j = 1, 2, \ldots, c-1$. Zgodnie z tym ujęciem formułę dotyczącą prawdopodobieństwa rozkładu oczekiwanego możemy przedstawić nastę-

pująco:

$$\pi_{ij}^{XY} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \xi^{(i-1)(j-1)} \cdot \phi_i^{(j-1)}, \tag{2.27}$$

zakładając, że kategoriami odniesienia są pierwsze kategorie obydwu zmiennych. W stosunku do 2.21 formuła zawiera parametr $\xi$. Można natomiast przyjąć, że dwa parametry $\phi_i$ — a nie jeden, jak zakładaliśmy przy formule 2.21 — są równe 1, np: $\phi_1 = \phi_r = 1$. Jest to równoznaczne z przyjęciem, że parametr $\delta$ w formule 2.26 wskazuje na stosunek szans wyznaczony dla skrajnych kategorii zmiennej $X$, a iloczyn parametrów $\delta_i$ z formuły jest równy 1.

Aby zinterpretować wielkości parametrów $\xi$ oraz $\phi_i$ z formuły 2.27 pomocne będzie wyznaczenie stosunku szans porównującego pierwsze i ostatnie kategorie obydwu zmiennych. Jak wiadomo, jest on równy iloczynowi wszystkich lokalnych stosunków szans i wynosi:

$$\Theta_{r/1;c/1}^{X\ Y} = \prod_{i=1}^{r-1}\prod_{i=1}^{c-1} \Theta_{ij}^{XY} = \xi^{(r-1)(c-1)}. \tag{2.28}$$

Pokazuje to, że w tym ujęciu parametr $\xi$ jest średnią geometryczną lokalnych stosunków szans wyznaczoną na podstawie rozkładu oczekiwanego zgodnego z hipotezą o efekcie wierszowym. W tym sensie jest on „odpowiednikiem" parametru z modelu jednakowej interakcji. Natomiast parametry wierszowe $\xi$ modyfikują jednakową interakcję poprzez wprowadzenie specyfiki zmiennej wierszowej. Parametr $\delta$ z formuły 2.26 można umownie interpretować jako średnią ogólną siły zależności. Ten sposób parametryzacji może być użyteczny dla omówienia pewnych zagadnień związanych z hipotezami dla trzech zmiennych, które przedstawimy w dalszej części tego rozdziału. Gdybyśmy założyli $\phi_1 = \phi_2 = 1$ wówczas interpretacja $\xi$ zmieniałaby się. Określał by on wówczas lokalne stosunki szans dla dwóch pierwszych kategorii zmiennej wierszowej.

W Tabeli 2.9 zamieszczony jest rozkład liczebności dwóch zmiennych (Europejski Sondaż Społeczny, 2004). Zmienna wierszowa $X$ opisuje położenie społeczno-zawodowe badanych. Wykorzystana została Społeczna Klasyfikacja Zawodów w wersji wyróżniającej 7 grup[6] Druga zmienna ($Y$) dotyczy zainteresowania polityką[7]. Okazuje się, że model wierszowy jest adekwatnym opisem danych: posiada 12 stopni swobody, statystyka $\chi^2 = 14,74$ ($p = 0,26$) natomiast statystyka $G^2 = 15,13$ ($p = 0,26$).

W tabeli 2.10 zostały przedstawione parametry modelu wierszowego. Kategorią odniesienia jest „inteligencja", dlatego dla tej kategorii wartość parametru jest równa

[6]Porównaj omówienie danych z tabeli 1.27.

[7]Na pytanie o zainteresowanie polityką jedynie trzy osoby odpowiedziały „Trudno powiedzieć". Zostały one wyłączone z analizy. Informacja o przynależności do grupy społeczno-zawodowej dotyczyła jedynie osób, które kiedykolwiek pracowały. Dane zostały przeważone zgodnie ze schematem doboru próby. W procesie weryfikacji przyjęto dla uproszczenia założenie o doborze prostym.

Tabela 2.9: Przynależność społeczno–zawodowa (klasyfikacja SKZ) a zainteresowanie polityką[a]

| Jak by Pan(i) określił(a) swoje zainteresowanie polityką? Czy Pan(i) polityką ... ($X$) | | | | |
|---|---|---|---|---|
| Kategorie społeczno–zawodowe ($Y$) | 1. bardzo się interesuje | 2. dosyć się interesuje | 3. niezbyt się interesuje | 4. w ogóle się nie interesuje |
| 1. Inteligencja | 12,6 | 52,4 | 30,2 | 4,1 |
| 2. Pozostali pracownicy umysłowi | 20,5 | 122,4 | 111,9 | 36,8 |
| 3. Pracownicy fizyczno–umysłowi | 10,2 | 67,1 | 84,4 | 37,8 |
| 4. Robotnicy wykwalifikowani | 28,1 | 124,9 | 162,0 | 72,4 |
| 5. Robotnicy nie wykwalifikowani | 8,0 | 42,7 | 95,5 | 44,7 |
| 6. Rolnicy | 7,0 | 43,4 | 71,8 | 41,4 |
| 7. Prywatni przedsiębiorcy | 5,9 | 69,9 | 56,0 | 16,4 |

[a]Źródło: Europejski Sondaż Społeczny, 2004. Dane przeważone.

1. Jeśli chodzi o zainteresowanie polityką, wartości parametrów wierszowych pokazują, że najbardziej zbliżeni do inteligencji są prywatni przedsiębiorcy ($\phi_2 = 1,59$) i pracownicy umysłowi ($\phi_2 = 1,61$). Przypomnijmy, że wartości parametrów wierszowych można interpretować jako stosunki szans dla sąsiednich kategorii zmiennej porządkowej $Y$ dla dwóch wybranych kategorii społeczno–zawodowych. Na przykład, zgodnie z hipotezą o efekcie wierszowym wśród prywatnych przedsiębiorców proporcja osób, które polityką *bardzo się interesują* w stosunku do osób, które *dość się interesują* jest 1,59 razy mniejsza niż wśród inteligencji (co określa iloraz $\phi_7/\phi_1$). Porównując do inteligencji inne grupy, stosunek ten jest jeszcze większy, na przykład wśród robotników

Tabela 2.10: Parametry modelu wierszowego $\phi_i$ dla danych z tabeli 2.9

| Kategoria społeczno–zawodowa | $\phi_i$ |
|---|---|
| 1. Inteligencja | 1,00 |
| 2. Pozostali pracownicy umysłowi | 1,61 |
| 3. Pracownicy fizyczno-umysłowi | 2,13 |
| 4. Robotnicy wykwalifikowani | 2,03 |
| 5. Robotnicy nie wykwalifikowani | 2,80 |
| 6. Rolnicy | 2,69 |
| 7. Prywatni przedsiębiorcy | 1,59 |

niewykwalifikowanych proporcja ta jest prawie 3 razy mniejsza. Można też zauważyć, że są grupy bardzo podobne do siebie pod względem zainteresowania polityką np. prywatni przedsiębiorcy i pozostali pracownicy umysłowi. Jak widać, $\phi_7/\phi_2$=0,99, czyli rozkłady warunkowe są dla tych dwóch grup prawie identyczne. Powyższe parametry pozwalają też porównywać rozkład zmiennej *zainteresowanie polityką* w poszczególnych grupach zawodowych. Rozkład tej zmiennej wśród inteligencji jest nie większy stochastycznie niż wśród prywatnych przedsiębiorców, a wśród tych ostatnich nie większy stochastycznie niż wśród rolników itd.

Jak zostało zasygnalizowane, choć zmienna wierszowa w modelu wierszowym (i odpowiednio zmienna kolumnowa w modelu kolumnowym) może być mierzona na skali nominalnej, to model ten może być również użyteczny w sytuacji, gdy obydwie zmienne są mierzone na skali porządkowej. Dzieje się tak wówczas, gdy związek pomiędzy zmiennymi jest bardziej skomplikowany aniżeli zakłada to hipoteza jednakowej interakcji. W tabeli 2.6 zostały zamieszczone wyniki weryfikacji hipotez o efekcie wierszowym (kolumnowym) w odniesieniu do danych z tabeli 2.5. Model wierszowy uwzględniający specyfikę poszczególnych kategorii zmiennej dotyczącej opinii w kwestii zmniejszenia zróżnicowania zarobków przez rząd, jest akceptowalny na poziomie istotności $\alpha = 0,01$, ale należałoby go odrzucić na poziomie istotności $\alpha = 0,05$.

Należy przypomnieć, że model jednakowej interakcji jest zagnieżdżony w modelu wierszowym. Porównując modele 2.9 i 2.21 widzimy, że $\phi_i = \delta^{i-1}$. Obydwa modele możemy porównać za pomocą testu warunkowego: $G^2 = 13,50 - 12,67 = 0,83$ przy 4 stopniach swobody, co pokazuje, że uwzględnienie specyfiki zmiennej wierszowej nie poprawia dopasowania do danych w sposób znaczący ($p = 0,93$). Parametry modelu wierszowego wynoszą odpowiednio: $\phi_1 = 1$, $\phi_2 = 0,69$, $\phi_3 = 0,42$, $\phi_4 = 0,42$, $\phi_5 = 0,28$. Gdybyśmy próbowali „zrekonstruować" powyższe wartości na podstawie kolejnych potęg parametru $\delta$ z modelu jednakowej interakcji wynosiłyby one odpowiednio: $\delta^0 = 1$, $\delta^1 = 0,74$, $\delta^2 = 0,54$, $\delta^3 = 0,39$, $\delta^4 = 0,29$. Różnice te są niewielkie, co potwierdza, że wprowadzenie oddzielnych parametrów dla zmiennej $X$, nie daje lepszego opisu niż znacznie prostszy model jednakowej interakcji.

Inaczej jest z modelem kolumnowym, który uwzględnia specyfikę zmiennej opisującej zadowolenie z sytuacji materialnej. Obydwie statystyki pokazują dobre dopasowanie do danych: model ten posiada 6 stopni swobody $\chi^2 = 7,12$ ($p = 0,31$), $G^2 = 7,19$ ($p = 0,30$) . Model ten posiada tylko 1 niezależny parametr więcej niż model jednakowej interakcji. Test warunkowy porównujący obydwa modele pokazuje, że $G^2 = 13,50 - 7,19 = 6,31$, ($p = 0,012$). Na poziomie istotności $\alpha = 0,01$, należałoby przyjąć model prostszy, jednak na poziomie $\alpha = 0,05$ dopasowanie modelu kolumnowego okazuje się istotnie statystycznie lepsze. Parametry $\phi_j$w modelu kolum-

Tabela 2.11: Rozkład oczekiwany zgodny z modelem kolumnowym dla danych z tabeli 2.5

| Rząd powinien zmniejszyć różnice w dochodach ($X$) | Zadowolenie z własnej sytuacji finansowej ($Y$) | | |
|---|---|---|---|
| | 1. Zadowolony(a) | 3. Mniej więcej zadowolony(a) | 3. Niezadowolony(a) 3 |
| 1. Zdecydowanie się zgadzam | 49,7 | 170,2 | 360,4 |
| 2. Zgadzam się | 58,6 | 180,8 | 244,7 |
| 3. Ani się zgadzam, ani nie zgadzam | 15,2 | 42,5 | 36,7 |
| 4. Nie zgadzam się | 12,6 | 31,7 | 17,5 |
| 5. Zdecydowanie się nie zgadzam | 4,1 | 9,2 | 3,2 |

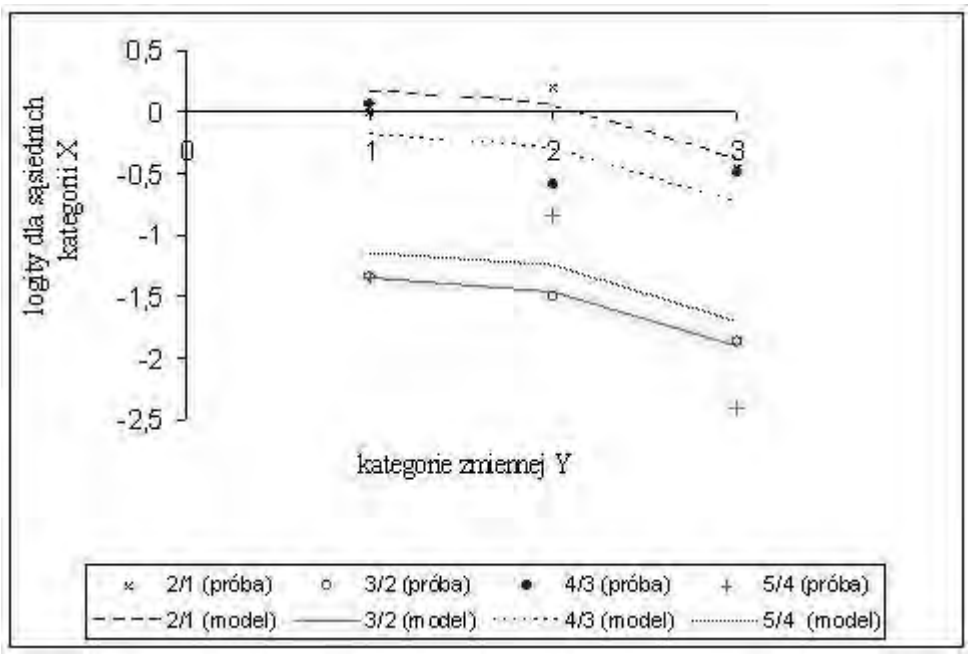
nowym są równe odpowiednio $\phi_1 = 1$, $\phi_2 = 0,90$, $\phi_3 = 0,57$, co pokazuje, że osoby *mniej więcej zadowolone* z sytuacji w opiniach zmniejszania różnic w dochodach różnią się nieznacznie w stosunku do osób *zadowolonych* natomiast różnią się znacznie w stosunku do osób *niezadowolonych* (kolejne potęgi parametru jednakowej interakcji $\delta$ wynoszą odpowiednio: $\delta^0 = 1$, $\delta^1 = 0,74$, $\delta^2 = 0,54$).

Tabela 2.11 pokazuje rozkład oczekiwany zgodny z modelem kolumnowym. Wykresy 2.3 i 2.4 pokazują graficzną interpretację modelu kolumnowego w odniesieniu do danych z tabeli 2.5, tj. odpowiednio logity warunkowe dla kategorii $X$ względem $Y$ na pierwszym wykresie, a na kolejnym logity warunkowe dla kategorii $Y$ względem $X$. Wybrany został model kolumnowy, gdyż jego dopasowanie do danych było lepsze, choć oczywiście analogicznie można zilustrować model wierszowy. Na rysunku 2.3 widać, że warunkowe logity dla sąsiednich kategorii zmiennej $X$ względem $Y$ nie leżą na jednej linii. Odległości pomiędzy logitami pomiędzy kolejnymi kategoriami zmiennej $Y$ wynoszą

$$\log\left(\Omega^{X\ (Y)}_{i+1/i(j+1)}\right) - \log\left(\Omega^{X\ (Y)}_{i+1/i(j)}\right) = \log\left(\frac{\Omega^{X\ (Y)}_{i+1/i(j+1)}}{\Omega^{X\ (Y)}_{i+1/i(j)}}\right) = \log\left(\frac{\phi_{j+1}}{\phi_j}\right) = b_{j+1} - b_j,$$

gdzie $b_j$, $b_{j+1}$ są parametrem efektu kolumnowego w wersji addytywnej. Wykresy dla logitów $\log\left(\Omega^{X\ (Y)}_{2/1(j+1)}\right)$, $\log\left(\Omega^{X\ (Y)}_{3/2(j+1)}\right)$, $\log\left(\Omega^{X\ (Y)}_{4/3(j+1)}\right)$, $\log\left(\Omega^{X\ (Y)}_{5/4(j+1)}\right)$ pozostają — podobnie jak w przypadku modelu jednakowej interakcji — równoległe do siebie. Na rysunku 2.4 na odwrót: widać, że warunkowe logity dla sąsiednich kategorii zmiennej

Rysunek 2.3: Warunkowe logity dla sąsiednich kategorii zmiennej $X$ względem $Y$ (model kolumnowy)




$Y$ względem $X$ leżą na jednej linii, jednak linie te nie są równoległe do siebie. Wynika to z tego, że wielkość parametrów kolumnowych określa kąt nachylenia linii.

Warto zauważyć, że w sytuacji, gdy weryfikujemy hipotezę o efekcie wierszowym (kolumnowym) a obydwie zmienne mierzone są na skali porządkowej, możemy zakładać — choć nie jest to konieczne — że związek pomiędzy zmiennymi będzie monotoniczny. Wyrażając to inaczej: można oczekiwać, że parametry wierszowe dla kolejnych kategorii zmiennej wierszowej będą niemalejące, bądź nierosnące (tak jak w analizach dotyczących danych z tabeli 2.5). Formalnie można zakładać, że zachodzi:

$$\begin{aligned} \phi_i &\leq \phi_{i+1}, \text{bądź} \\ \phi_i &\geq \phi_{i+1}, \text{dla każdego } i < r-1. \end{aligned} \tag{2.29}$$

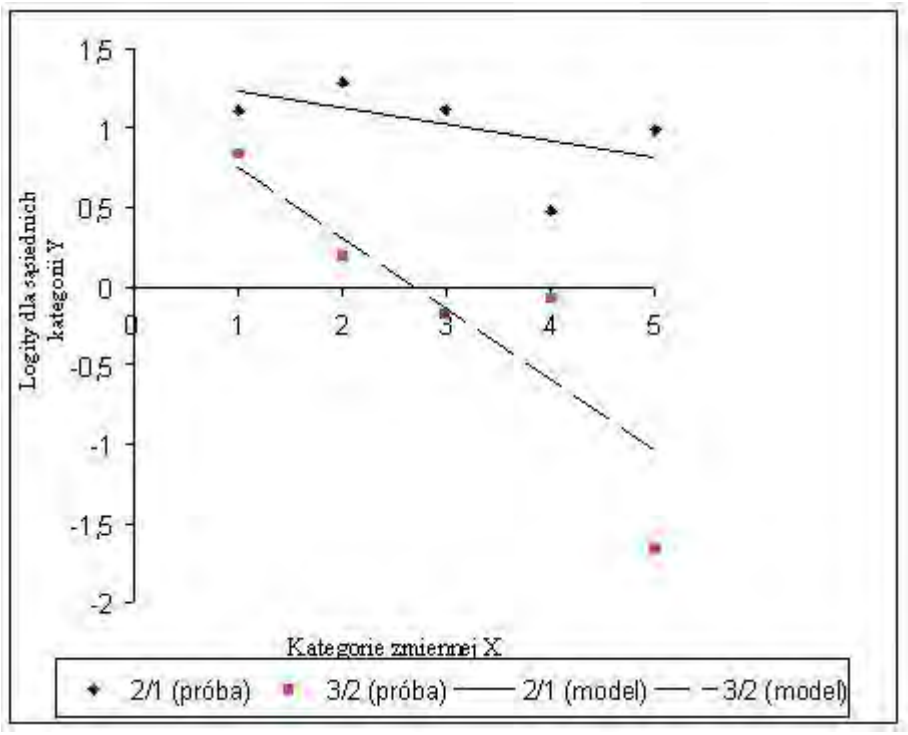
Nałożenie takiego warunku jest możliwe i często wydaje się sensowne. Jeśli na przykład zachodzi

$$\phi_1 < \phi_2 < \phi_3 > \phi_4,$$

można przypuszczać, że zakłócenie monotoniczności (relacja pomiędzy $\phi_3$ oraz $\phi_4$) jest przypadkowa tj. może wynikać z tego, że posługujemy się próbą losową. W praktyce dodanie założenia $\phi_3 \leq \phi_4$ zgodnie estymacją największej wiarygodności prowadzi do wyniku $\phi_3 = \phi_4$[8]. W tym miejscu sygnalizujemy jedynie możliwość wprowadzenia

[8]Wynika, to z tego, że funkcja wiarygodności jest wypukła

Rysunek 2.4: Warunkowe logity dla sąsiednich kategorii zmiennej $Y$ względem $X$ (model kolumnowy)




takiego warunku, problem ten bardziej szczegółowo został omówiony w artykułach Goodmana (1985) oraz Agrestiego i innych (1987).

### 2.3.3 Model efektu wierszowo–kolumnowego (RC1)

Model wierszowy (bądź kolumnowy) może opisywać sytuację, gdy jedna zmienna mierzona jest na skali porządkowej a druga na skali nominalnej. W niektórych sytuacjach model ten może również opisywać związek pomiędzy dwiema zmiennymi porządkowymi bardziej adekwatnie niż model jednakowej interakcji, tj. interakcja pomiędzy zmiennymi jest specyficzna dla każdej pary wierszy. Możliwe jest równoczesne uwzględnienie specyfiki poszczególnych kategorii obydwu zmiennych porządkowych. Model taki nazywamy modelem *wierszowo–kolumnowym*, będzie on oznaczany jak (RC1 — *row–column model I*)[9]. Lokalny stosunek szans w tym modelu zależy z jednej strony od wartości zmiennej $X$, z drugiej strony od wartości zmiennej $Y$. Zapiszmy to jako:

$$\Theta_{ij}^{XY} = \delta_{i\cdot} \cdot \delta_{\cdot j} \tag{2.30}$$

[9]W literaturze znane są dwa modele nazywane wierszowo-kolumnowymi. Oznaczane są odpowiednio jako pierwszy i drugi. Drugi z tych modeli — zwany również modelem logarytmiczno–multiplikatywnym — omówiony zostanie z następnym paragrafie

Nie oznacza to jednak, że stosunek szans jest specyficzny dla każdej kombinacji obydwu zmiennych tak jak ma to miejsce w modelu nasyconym. Można pokazać, że iloraz dwóch stosunków lokalnych szans $\Theta_{aj}^{XY}$, $\Theta_{bj}^{XY}$ wyróżnionych dla tej samej kategorii $y_j$ zmiennej kolumnowej $Y$ i dwóch wartości $x_a$, $x_b$ zmiennej wierszowej $X$ jest taki sam dla każdej kategorii zmiennej $Y$, tj:

$$\frac{\Theta_{aj}^{XY}}{\Theta_{bj}^{XY}} = \frac{\delta_{a\cdot} \cdot \delta_{\cdot j}}{\delta_{b\cdot} \cdot \delta_{\cdot j}} = \frac{\delta_{a\cdot}}{\delta_{b\cdot}} \tag{2.31}$$

Jak widać wielkość ta nie zależy od tego, dla której kategorii zmiennej $Y$ wyróżniono obydwa lokalne stosunki szans a jedynie od tego, z którymi kategoriami zmiennej $X$ mamy do czynienia. Analogicznie daje się pokazać, że:

$$\frac{\Theta_{ic}^{XY}}{\Theta_{id}^{XY}} = \frac{\delta_{i\cdot} \cdot \delta_{\cdot c}}{\delta_{i\cdot} \cdot \delta_{\cdot d}} = \frac{\delta_{\cdot c}}{\delta_{\cdot d}} \tag{2.32}$$

Tabela 2.12: Ilustracja modelu wierszowo–kolumnowego I typu — parametry interakcji

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\psi_2 \cdot \phi_2$ | $\psi_2^2 \cdot \phi_3$ | $\psi_2^3 \cdot \phi_4$ |
| $x_3$ | 1 | $\psi_3 \cdot \phi_2^2$ | $\psi_3^2 \cdot \phi_3^2$ | $\psi_3^3 \cdot \phi_4^2$ |
| $x_4$ | 1 | $\psi_4 \cdot \phi_2^3$ | $\psi_4^2 \cdot \phi_3^3$ | $\psi_4^3 \cdot \phi_4^3$ |

Można zresztą zauważyć, że warunek 2.32 wynika z warunku 2.31. Ilustrację modelu stanowi tabela 2.12, przy czym wielkości w niej przedstawione zostaną zdefiniowane w dalszej części. Zgodnie z powyższymi warunkami zachodzi na przykład:

$$\frac{\Theta_{21}^{XY}}{\Theta_{31}^{XY}} = \frac{\Theta_{22}^{XY}}{\Theta_{32}^{XY}} = \frac{\Theta_{23}^{XY}}{\Theta_{33}^{XY}},$$

jak również,

$$\frac{\Theta_{11}^{XY}}{\Theta_{12}^{XY}} = \frac{\Theta_{21}^{XY}}{\Theta_{22}^{XY}} = \frac{\Theta_{31}^{XY}}{\Theta_{32}^{XY}}.$$

Porównując model wierszowo–kolumnowy z modelami omawianymi wcześniej, można zauważyć, że zdefiniowane powyżej ilorazy 2.31, 2.32 są równe 1 w modelu jednakowej interakcji. W modelu wierszowym iloraz 2.31 jest równy $\delta_i$, dla każdej kategorii $y_j$ (przy założeniu, że $X$ jest zmienną wierszową), a iloraz 2.32 jest równy 1. W modelu kolumnowym iloraz 2.31 jest równy 1, a iloraz 2.32 jest równy $\delta_j$. Każdy z tych modeli — jednakowej interakcji, wierszowy i kolumnowy — jest prostszy od omawianego modelu wierszowo–kolumnowego. Nakładając na model wierszowo–kolumnowy warunek

$\delta_{\cdot j} = const$ otrzymujemy model wierszowy. Nałożenie warunku $\delta_{i\cdot} = const$ prowadzi do sformułowania modelu kolumnowego. Obydwa warunki nałożone jednocześnie dają model jednakowej interakcji.

Związek z niemniejszością stochastyczną rozkładów warunkowych jest bardziej skomplikowany niż w przypadku modeli prezentowanych wcześniej. Jeśli wielkości $\delta_{\cdot j}$ są rosnące względem kategorii zmiennej porządkowej $Y$ tj. zachodzi $\delta_{\cdot(j+1)} > \delta_{\cdot j}$ wówczas rozkład $Y$ w podzbiorowości $x_a$ jest nie mniejszy stochastycznie niż w podzbiorowości $x_b$ jeśli $\delta_{a\cdot} > \delta_{b\cdot}$. Analogiczne warunki implikują niemniejszość stochastyczną rozkładów $X$ względem $Y$.

Powyższa tabela 2.12 jest zgodna z parametryzacją względem kategorii odniesienia, przy czym są nimi pierwsze kategorie zmiennych $X$ oraz $Y$. Wyrażenia $\psi_i^{(j-1)} \cdot \phi_j^{(i-1)}$ definiują stosunki szans typu $\Theta^{X\ Y}_{i/1;j/1}$. Jeśli kategoriami odniesienia są kategorie $x_a$ oraz $y_b$ wówczas model można zapisać jako:

$$\pi_{ij}^{XY} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \psi_i^{(j-a)} \cdot \phi_j^{(i-b)}. \tag{2.33}$$

Przy czym zachodzą warunki 1.53 oraz $\psi_1$=1 oraz $\phi_1$=1. Dla parametryzacji odchyleń multiplikatywnych powyższy model przyjmuje postać:

$$\pi_{ij}^{XY} = \tau \cdot \tau_i^X \cdot \tau_j^Y \cdot \psi_i^{(j-\bar{j})} \cdot \phi_j^{(i-\bar{i})}. \tag{2.34}$$

Podobnie jak w poprzednich modelach na parametry $\tau_i^X$, $\tau_j^Y$ nakładamy warunki 1.36, $\bar{i}$ oraz $\bar{j}$ są takimi stałymi, że, $\sum(i-\bar{i}) = 0$ oraz $\sum(j-\bar{j}) = 0$ a iloczyny wszystkich parametrów $\psi_i$ oraz $\phi_j$ są równe 1, tj:

$$\prod_{i=1}^{r} \psi_i = 1 \qquad \text{oraz} \qquad \prod_{j=1}^{c} \phi_j = 1 \tag{2.35}$$

W formie addytywnej model wierszowo-kolumnowy można przedstawić jako:

$$\log \pi_{ij}^{XY} = \mu + \lambda_i^X + \lambda_j^Y + b_{i\cdot}(j-\bar{j}) + b_{\cdot j}(i-\bar{i}) \tag{2.36}$$

Aby móc jednoznacznie oszacować metodą największej wiarygodności rozkład oczekiwany zgodny z hipotezą o efekcie wierszowo–kolumnowym, poza warunkami 2.13, 2.14 muszą zachodzić równocześnie warunki podobne jak w przypadku modelu wierszowego i modelu kolumnowego:

$$\sum_{i=1}^{r} i \cdot \widehat{\pi}_{ij}^{XY} = \sum_{i=1}^{r} i \cdot p_{ij}^{XY}, \tag{2.37}$$

$$\sum_{j=1}^{c} j \cdot \widehat{\pi}_{ij}^{XY} = \sum_{j=1}^{c} j \cdot p_{ij}^{XY}, \tag{2.38}$$

dla każdego $i$ oraz każdego $j$, gdzie $p_{ij}$ to częstości z próby. Podobnie jak w modelach jednakowej interakcji, wierszowego i kolumnowego nie istnieje formuła pozwalająca na skonstruowanie rozkładu oczekiwanego i konieczne jest zastosowanie metod iteracyjnych. Zanim omówimy liczbę niezależnych parametrów modelu wierszowo–kolumnowego przedstawimy go w alternatywnej parametryzacji. Będzie ona analogiczna do do formuły 2.26, która dotyczyła hipotezy o efekcie wierszowym. Hipotezę 2.30 można sformułować równoważnie:

$$\Theta_{ij}^{XY} = \delta \cdot \delta_{i\cdot} \cdot \delta_{\cdot j} \tag{2.39}$$

Zakładając, że kategoriami odniesienia są pierwsze kategorie obydwu zmiennych formułę na prawdopodobieństwa oczekiwane można przedstawić jako:

$$\pi_{ij}^{XY} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \xi^{(i-1)(j-1)} \cdot \psi_i^{(j-1)} \cdot \phi_j^{(i-1)} \tag{2.40}$$

Można przyjąć, że dwa parametry wierszowe i dwa modele kolumnowe są równe 1, na przykład $\psi_1 = \psi_r = \phi_1 = \phi_r = 1$. Wówczas parametr $\xi$ — analogicznie jak w formule 2.28 — może być interpretowany jako średnia geometryczna lokalnych stosunków szans wyznaczonych na podstawie rozkładu oczekiwanego dla hipotezy wierszowo–kolumnowej. Parametry $\psi_i$, $\phi_j$ opisują wierszowe i kolumnowe modyfikacje tej zależności. Tabela 2.13 pokazuje lokalne stosunki szans wyrażone w parametrach równania 2.40.

Tabela 2.13: Lokalne stosunki szans zgodne z modelem wierszowo–kolumnowym 2.40

| $i\backslash j$ | 1 | 2 | 3 |
|---|---|---|---|
| 1 | $\xi\psi_2\phi_2$ | $\xi\psi_2\frac{\phi_3}{\phi_2}$ | $\xi\psi_2\frac{1}{\phi_3}$ |
| 2 | $\xi\frac{\psi_3}{\psi_2}\phi_2$ | $\xi\frac{\psi_3}{\psi_2}\frac{\phi_3}{\phi_2}$ | $\xi\frac{\psi_3}{\psi_2}\frac{1}{\phi_3}$ |
| 3 | $\xi\frac{1}{\psi_3}\phi_2$ | $\xi\frac{1}{\psi_3}\frac{\phi_3}{\phi_2}$ | $\xi\frac{1}{\psi_3}\frac{1}{\phi_3}$ |

Na przecięciu $i$–tego wiersza oraz $j$–tej kolumny mamy podaną formułę dotyczącą lokalnego stosunku szans $\Theta_{ij}^{XY}$. Prześledzenie tych formuł pokazuje, że przyjęcie takich założeń nie wpływa na zmianę ogólności modelu. Widać na przykład, że spełnione są warunki 2.31, 2.32 określające równość odpowiednich ilorazów stosunków szans.

Takie ujęcie modelu wierszowo–kolumnowego pokazuje, że posiada on $(r-2)+(c-2)-1$ niezależnych parametrów więcej aniżeli model niezależności stochastycznej. Liczba stopni swobody modelu wierszowo–kolumnowego wynosi:

$$df = (r-1)(c-1) - (r-2) - (c-2) - 1 = (r-2)(c-2).$$

Liczbę stopni swobody można wyrazić również odwołując się do warunków 2.31 lub 2.32 odnoszących się do ilorazów stosunków szans. Dla każdej pary sąsiednich kategorii zmiennej kolumnowej można wyznaczyć $(r-2)$ ilorazów opisanych w powyższym warunku, które są niezależne od siebie. Odpowiednie ilorazy są równe dla kolejnych kolumn, co przekłada się łącznie na $(r-2)(c-2)$ warunków.

Tabela 2.6 pokazuje, że model wierszowo-kolumnowy (RC1) stanowi akceptowalny opis danych z tabeli 2.5. Statystyka $G^2 = 6,69$ przy 3 stopniach swobody ($p = 0,50$). Parametry wierszowe dla tego modelu wynoszą odpowiednio $\psi_1 = 1$, $\psi_2 = 0,803$, $\psi_3 = 0,692$, $\psi_4 = 0,609$, $\psi_5 = 0,4$ a parametry kolumnowe $\phi_1 = 1$, $\phi_2 = 1,090$, $\phi_3 = 1,084$. W tabeli 2.14 zostały podane liczebności oczekiwane zgodne z omawianym modelem. Aby lepiej zrozumieć specyfikę tego modelu porównajmy wartości wybranych lokalnych stosunków szans, wyrażając je z jednej strony za pomocą parametrów wierszowych i kolumnowych, z drugiej strony za pomocą liczebności:

$$\Theta_{11}^{XY} = \frac{\psi_2 \cdot \phi_2}{\psi_1 \cdot \phi_1} = \frac{49,0 \cdot 182,3}{169,3 \cdot 60,2} = 0,87$$

$$\Theta_{12}^{XY} = \frac{\psi_2 \cdot \phi_3}{\psi_1 \cdot \phi_2} = \frac{169,3 \cdot 241,6}{362,1 \cdot 182,3} = 0,62$$

Tabela 2.14: Rozkład oczekiwany zgodny z modelem wierszowo-kolumnowym dla danych z tabeli 2.5

| Rząd powinien zmniejszyć różnice w dochodach ($X$) | Zadowolenie z własnej sytuacji finansowej ($Y$) | | |
|---|---|---|---|
| | 1. Zadowolony(a) | 3. Mniej więcej zadowolony(a) | 3. Niezadowolony(a) 3 |
| 1. Zdecydowanie się zgadzam | 49,0 | 169,3 | 362,1 |
| 2. Zgadzam się | 60,2 | 182,3 | 241,6 |
| 3. Ani się zgadzam, ani nie zgadzam | 14,9 | 42,3 | 37,3 |
| 4. Nie zgadzam się | 11,5 | 31,5 | 18,8 |
| 5. Zdecydowanie się nie zgadzam | 4,7 | 9,1 | 2,8 |

Jak widać, lokalne stosunki szans uwzględniają zarówno parametry wierszowe jak też kolumnowe. Zgodnie z modelem jednakowej interakcji proporcja liczby osób, które *zdecydowanie zgadzają się* co do tego, że rząd powinien zmniejszyć różnice w dochodach ($X = x_1$) do liczby respondentów, którzy *zgadzają się* z taką opinią ($X = x_2$)

jest 1,14 (1/0,87) razy mniejsza, wśród tych, którzy są *zadowoleni* z sytuacji finansowej ($Y = y_1$) aniżeli wśród osób *średnio zadowolonych* ($Y = y_2$). Proporcja ta, jest za to 1,61 (1/0,62) razy większa wśród osób *niezadowolonych* z własnych dochodów, aniżeli wśród osób *zadowolonych*. Porównanie tych dwóch wielkości daje nam iloraz

$$\Theta_{11}^{XY}/\Theta_{12}^{XY} = \phi_2/\phi_1 = 0,87/0,62 = 1,41.$$

Jeśli rozpatruje się powyższą proporcję dla innej pary kategorii zmiennej $X$, na przykład osób *zgadzających się* z opinią dotyczących dochodów i *neutralnych* wobec niej, iloraz ten byłby — zgodnie z (2.31) — taki sam:

$$\Theta_{21}^{XY}/\Theta_{22}^{XY} = 0,94/0,66 = 1,41.$$

Przypomnijmy, że omawiany model jest zagnieżdżony zarówno w modelu jednakowej interakcji, jak również w modelu wierszowym i w modelu kolumnowym. Testy warunkowe pokazują, że w odniesieniu do modelu jednakowej interakcji i modelu kolumnowego model ten nie poprawia dopasowania w sposób znaczący. Odpowiednie statystyki wynoszą: $G^2 = 13,5 - 6,69 = 6,81$ przy 4 stopniach swobody ($p = 0,15$) dla modelu jednakowej interakcji i $G^2 = 7,19 - 6,69 = 0,5$ przy 3 stopniach swobody ($p = 0,92$) dla modelu kolumnowego. Natomiast model wierszowy na poziomie istotności $\alpha = 0,05$ należałoby odrzucić na rzecz modelu wierszowo–kolumnowego: $G^2 = 12,67 - 6,69 = 5,98$ przy 1 stopniu swobody ($p = 0,014$).

### 2.3.4 Logarytmiczno–multiplikatywny model wierszowo–kolumnowy (RC2)

Jak zostało powiedziane wcześniej model jednakowej interakcji zakłada, że wszystkie lokalne stosunki szans są sobie równe. W pewnych sytuacjach można oczekiwać, że wartości stosunków szans zależą od tego na ile „podobne" są do siebie sąsiadujące kategorie zmiennej porządkowej. Przypuśćmy, ze posiadamy informacje o wykształceniu zakodowane w następujący sposób: 1. niepełne podstawowe, 2. podstawowe, 3. nieukończone średnie, 4. średnie 5. wyższe. O ile możemy się zgodzić do porządkowego charakteru tej zmiennej to niekoniecznie kolejne szczeble wykształcenia muszą wiązać się z takim samym „przyrostem" wiedzy, czy też wymagań dotyczących kompetencji i umiejętności. Może to znajdować odzwierciedlenie, jeśli badamy związek wykształcenia z innymi cechami badanych osób, np. opiniami dotyczącymi kwestii gospodarczych. W niektórych sytuacjach można oczekiwać, że poglądy osób posiadających wykształcenie *niepełne podstawowe* i *podstawowe* mogą być do siebie bardziej zbliżone niż poglądy osób posiadających wykształcenie *średnie* i *wyższe*, pomimo, że w jednym i w drugim przypadku wymienione kategorie sąsiadują ze sobą.

Problem ten może pojawiać się dość często w analizie rozkładu łącznego. W modelu wierszowo–kolumnowym II typu próbujemy uwzględnić odległości pomiędzy kategoriami obydwu zmiennych. Aby to uczynić do kolejnych wartości zmiennej $X$ przypisuje się oceny $u_1$, $u_2$, $u_3$,..., $u_i$,... $u_r$, które mają odzwierciedlać odległości pomiędzy kategoriami zmiennej, np. różnica $u_{(i+1)} - u_i$ ma odzwierciedlać różnice pomiędzy sąsiednimi kategoriami zmiennej $X$, tj. kategoriami $x_{(i+1)}$ oraz $x_i$. Analogicznie do kategorii zmiennej $Y$ przypisuje się oceny $v_1$, $v_2$, $v_3$, ..., $v_j$,... $v_c$. Hipoteza związana z drugą wersja modelu wierszowo–kolumnowego głosi, że wielkość lokalnego stosunku szans uwzględnia odległości pomiędzy kategoriami zmiennej, tj.

$$\Theta_{ij}^{XY} = \delta^{(u_{(i+1)}-u_i)(v_{(i+1)}-v_i)} \text{ dla każdej pary } i,\ j, \text{ takich, że } i \leq r-1,\ j \leq c-1. \tag{2.41}$$

Należy zauważyć, że jeśli odległości pomiędzy kolejnymi wartościami zmiennej $X$ i zmiennej $Y$ są takie same tj. $u_{(i+1)} - u_i = const$ oraz $v_{(j+1)} - v_j = const$ wówczas warunek 2.41 byłby zbieżny z hipotezą o jednakowej interakcji. Model wierszowo–kolumnowy modyfikuje wielkości lokalnych stosunków szans biorąc pod uwagę, że odległości pomiędzy sąsiednimi kategoriami zmiennej mogą być mniejsze bądź większe. Wielkości $u_i$, $v_j$ są parametrami, które szacuje się w modelu.

Jeśli oszacowane parametry są monotoniczne względem kategorii zmiennej $X$ oraz zmiennej $Y$, daje się pokazać, zależność pomiędzy omawianą hipotezą a niemniejszością stochastyczną. Jeśli np. zachodzi:

$$u_1 < u_2 < u_3 < \ldots < u_i < \ldots < u_r, \tag{2.42}$$

wówczas rozkłady warunkowe $X$ pozostają w relacji niemniejszości stochastycznej, tj. rozkład warunkowy $X$ jest niemniejszy stochastycznie w podzbiorowości $y_b$ niż w podzbiorowości $y_a$ o ile $v_b < v_a$. Podobną zależność odnośnie rozkładów warunkowych $Y$ można pokazać, jeśli spełniony jest warunek:

$$v_1 < v_2 < v_3 < \ldots < v_i < \ldots < v_c. \tag{2.43}$$

Jeśli spełnione są obydwa warunki, zmienne pozostają w relacji zależności regresyjnej. Tak jak w przypadku modeli omawianych poprzednio związek ten nie zachodzi w druga stronę: niemniejszość stochastyczna, czy nawet, zależność regresyjna nie implikuje tego, że spełniona będzie hipoteza o efekcie wierszowo–kolumnowym II typu.

Ponieważ wielkości $u_i$, $v_j$ są parametrami szacowanymi w modelu, daje to pewien rodzaj skalowania wartości zmiennej porządkowej. Warto pamiętać, ze skalowanie to zawsze odbywa się względem drugiej zmiennej występującej w modelu — nazywanej często zmienną *instrumentalną* — a dokładniej na podstawie związku pomiędzy tymi zmiennymi. Skalowanie jest więc z jednej strony zrelatywizowane do drugiej zmiennej i

jest na tyle trafne, na ile słuszne jest założenie o jednakowej interakcji zmodyfikowane przez uwzględnienie odległości między zmiennymi.

Tabela 2.15: Ilustracja modelu wierszowo–kolumnowego II typu — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\delta^{u_2v_2}$ | $\delta^{u_2v_3}$ | $\delta^{u_2v_4}$ |
| $x_3$ | 1 | $\delta^{u_3v_2}$ | $\delta^{u_3v_3}$ | $\delta^{u_3v_4}$ |
| $x_4$ | 1 | $\delta^{u_4v_2}$ | $\delta^{u_4v_3}$ | $\delta^{u_4v_4}$ |

Tabela 2.15 ilustruje związek między zmiennymi w modelu wierszowo–kolumnowym. Tak jak poprzednio ilustracja ta jest zgodna z parametryzacją względem kategorii odniesienia, którymi są pierwsze kategorie obydwu zmiennych. Bardziej ogólnie, jeśli kategoriami odniesienia są kategorie $x_a$ oraz $y_b$ wówczas model można zapisać jako:

$$\pi_{ij}^{XY} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \delta^{(u_i-u_a)(v_i-v_b)}. \tag{2.44}$$

Przy czym zachodzą warunki 1.53. Przyjmujemy wówczas, że $u_a = 0$, oraz $v_b = 0$. Ponieważ interesują nas nie tyle bezwzględne wielkości parametrów, co odległości pomiędzy nimi, możliwe jest nałożenie na ich wartości dodatkowych warunków, np:

$$\sum_{i=1}^{r} u_i^2 = \sum_{j=1}^{c} v_j^2 = 1, \tag{2.45}$$

Warunki te określają „jednostkę" ocen przypisanych do poszczególnych wartości. Możliwe jest przyjęcie innych warunków (Goodman 1979*b*, Agresti 1984). Dla parametryzacji odchyleń multiplikatywnych powyższy model przyjmuje postać:

$$\pi_{ij}^{XY} = \tau \cdot \tau_i^X \cdot \tau_j^Y \cdot \delta^{(u_i-\bar{u}_i)(v_i-\bar{v}_j)}. \tag{2.46}$$

Podobnie jak w poprzednich modelach na parametry $\tau_i^X$, $\tau_j^Y$ nakładamy warunki 1.36, natomiast na wielkości $u_i$ oraz $v_j$:

$$\sum_{i=1}^{r} u_i = \sum_{j=1}^{c} v_j = 0 \qquad \text{oraz} \qquad \sum_{i=1}^{r} u_i^2 = \sum_{j=1}^{c} v_j^2 = 1. \tag{2.47}$$

Wielkości $\bar{u}_i$ oraz $\bar{u}_j$ są takimi stałymi, że $\sum_i (u_i - \bar{u}_i) = 0$ oraz $\sum_j (v_j - \bar{v}_j) = 0$. Addytywną formę tego modelu można zapisać jako:

$$\log \pi_{ij}^{XY} = \mu + \lambda_i^X + \lambda_j^Y + b(u_i - \bar{u}_i)(v_j - \bar{v}_j). \tag{2.48}$$

Pomimo, że mamy do czynienia z formą addytywną parametry $b$, $u_i$, $v_j$ są mnożone przez siebie. Dlatego też model wierszowo–kolumnowy typu II nie może być postrzegany jako szczególny przypadek uogólnionego modelu liniowego. W istocie model ten nie jest modelem logarytmiczno–liniowym, stąd też nazywa się go modelem *logarytmiczno–multiplikatywnym*

Aby oszacować metodą największej wiarygodności rozkład oczekiwany zgodny z hipotezą o efekcie wierszowo–kolumnowym, poza warunkami 2.13, 2.14 muszą zachodzić warunki analogiczne jak w przypadku modelu wierszowego–kolumnowego typu I:

$$\sum_{i=1}^{r} \widehat{u}_i \cdot \widehat{\pi}_{ij}^{XY} = \sum_{i=1}^{r} \widehat{u}_i \cdot p_{ij}^{XY}, \tag{2.49}$$

$$\sum_{j=1}^{c} \widehat{v}_j \cdot \widehat{\pi}_{ij}^{XY} = \sum_{j=1}^{c} \widehat{v}_j \cdot p_{ij}^{XY}, \tag{2.50}$$

dla każdego $i$, $j$, gdzie $p_{ij}$ to częstości z próby. Podobnie jak w modelach jednakowej interakcji, wierszowym i kolumnowym nie istnieje formuła pozwalająca na skonstruowanie rozkładu oczekiwanego i konieczne jest zastosowanie metod iteracyjnych (Goodman 1979*b*). W procedurach tych, zazwyczaj przyjmuje się arbitralnie pewne wielkości początkowe parametrów $\widehat{u}_i$, i wobec nich oszacowuje się wielkości drugiej zmiennej jako $\widehat{v}_j^*$. W kolejnym kroku na odwrót: względem oszacowanych w ten sposób wielkości $\widehat{v}_j^*$ oszacowuje się parametry drugiej zmiennej jako $\widehat{u}_i^*$. Procedurę tę powtarza się aż do uzyskania zbieżności, tj. kolejne kroki nie zmieniają oszacowań $\widehat{u}_i^*$, $\widehat{v}_j^*$.

Omawiany model jako logarytmiczno–multiplikatywny posiada pewne ograniczenia w porównaniu do modeli logarytmiczno–liniowych. Mogą się pojawić pewne kłopoty przy estymacji tego modelu ponieważ funkcja wiarygodności związana z tym modelem nie jest wypukła. W konsekwencji procedury iteracyjne mogą prowadzić do rozwiązania, które jest lokalnym maksimum tej funkcji (tj. możliwe jest znalezienie rozwiązania lepszego). Ponadto, ograniczone są możliwości porównywania tego modelu za pomocą testów warunkowych. Na przykład różnica w dopasowaniu modelu $G^2(RCII)$ i modelu niezależności $G^2(I)$ nie musi dążyć asymptotycznie do rozkładu $\chi^2$ (Haberman 1981). Wynika to z tego, że z hipotezy o niezależności nie wynika nic odnośnie parametrów $u_i$, $v_j$, związanych z modelem logarytmiczno–multiplikatywnym[10].

Logarytmiczno–multiplikatywny model wierszowo–kolumnowy ma jeden parametr $\delta$, $(r-2)$ niezależnych parametrów $u_i$ oraz $(c-2)$ niezależnych parametrów $v_j$ więcej

[10]Istnieją alternatywne testy związane z analizą korelacji kanonicznej do testowania hipotezy głoszącej, że parametr interakcji $\gamma = 1$, lub $b = 0$ w postaci addytywnej (Haberman 1981). Ich omówienie przekracza ramy tej pracy.

w porównaniu do modelu niezależności stochastycznej. Wynika, to z warunków 2.47 ograniczających wielkości $u_i$, $v_j$. Liczba stopni swobody tego modelu wynosi:

$$df = (r-1)(c-1) - 1 - (r-2) - (c-2) = (r-2)(c-2).$$

Tabela 2.16: Rozkład oczekiwany zgodny z logarytmiczno–multiplikatywnym modelem wierszowo–kolumnowym dla danych z tabeli 2.5

| Rząd powinien zmniejszyć różnice w dochodach ($X$) | Zadowolenie z własnej sytuacji finansowej ($Y$) | | |
|---|---|---|---|
| | 1. Zadowolony(a) | 2. Mniej więcej zadowolony(a) | 3. Niezadowolony(a) 3 |
| 1. Zdecydowanie się zgadzam | 50,0 | 164,4 | 365,9 |
| 2. Zgadzam się | 61,0 | 187,0 | 236,1 |
| 3. Ani się zgadzam, ani nie zgadzam | 14,7 | 43,0 | 36,7 |
| 4. Nie zgadzam się | 10,3 | 29,7 | 21,8 |
| 5. Zdecydowanie się nie zgadzam | 4,2 | 10,3 | 2,0 |

Tabela 2.6 pokazuje, że logarytmiczno–multiplikatywny model wierszowo-kolumnowy (RC2) stanowi akceptowalny opis danych z tabeli 2.5. Statystyka $G^2$ wynosi 4,54 przy 3 stopniach swobody ($p = 0,20$). Indeks rozbieżności pokazuje, że niecałe 1,5 % osób w tabeli rozkładu łącznego jest zaklasyfikowanych niezgodnie z tym modelem. W tabeli 2.16 zamieszczone zostały liczebności oczekiwane związane z tym modelem. Interesujące jest prześledzenie wartości parametrów skalujących odległość pomiędzy kategoriami obydwu zmiennych. Są one równe odpowiednio: $u_1 = -0,564$, $u_2 = -0,247$, $u_3 = -0,028$, $u_4 = 0,054$, $u_5 = 0,786$ dla zmiennej $X$, oraz $v_1 = 0,47$, $v_2 = 0,336$, $v_3 = -0,812$, dla zmiennej $Y$. Oceny te są zgodne z uporządkowaniem kategorii obydwu zmiennych, warto jednak zauważyć, ze odległości pomiędzy kolejnymi kategoriami nie są takie same. Na przykład w odniesieniu do zmiennej opisującej opinie w sprawie zróżnicowania dochodów, odległość pomiędzy kategoriami *nie zgadzam się* oraz *ani się zgadzam, ani się nie zgadzam* jest niewielka (0,054-(-0,028)=0,082), natomiast pomiędzy kategoriami *nie zgadzam się* i *zdecydowanie się nie zgadzam* jest dużo większa (0,723). Podobnie daje się zauważyć, że w odniesieniu do sytuacji finansowej kategoria *mniej więcej zadowolony* jest znacznie bliższa kategorii *zadowolony* niż kategorii *niezadowolony*.

Parametr interakcji $\delta$ jest równy 1,56. Nie powinniśmy się jednak sugerować tą wielkością, która różni się znacznie od parametru jednakowej interakcji. To, że parametr ten jest większy od 1, wynika z tego, że oszacowane w modelu oceny dla zmiennej $Y$ są malejące względem kategorii zmiennej porządkowej. Możliwa jest zmiana znaków tych parametrów, wówczas, parametr interakcji wynosiłby 0,64 (1/1,56). Po drugie, należy pamiętać, że interpretacja tego parametru jest inna niż w modelu jednakowej interakcji: opisuje on lokalny stosunek szans — zgodnie z 2.41 — dopiero po uwzględnieniu parametrów skalujących wartości zmiennych, np.

$$\Theta_{11}^{XY} = 1,56^{(-0,247+0,564)(0,336-0,47)} = 0,98$$

Warto zauważyć, że parametry logarytmiczno–multiplikatywnego modelu wierszowo–kolumnowego nie muszą być monotoniczne względem kategorii zmiennej porządkowej, tj. dla $i = 1, 2...r-1$ nie musi zachodzić żadna z poniższych nierówności:

$$u_i \leq u_{i+1}, \text{ bądź } u_i \geq u_{i+1}. \tag{2.51}$$

Nałożenie takiego warunku wydaje się jednak pożądane w odniesieniu do zmiennej porządkowej. Możliwość przyjęcia podobnego warunku sygnalizowaliśmy w odniesieniu do modelu wierszowego (bądź kolumnowego). Przypomnijmy, że warunek 2.29 dotyczył hipotezy dotyczącej monotonicznego związku pomiędzy zmiennymi. W odniesieniu do omawianego modelu za przyjęciem analogicznego warunku 2.51 przemawiają dodatkowe argumenty. Jeśli parametry $u_i$ mają odzwierciedlać odległości pomiędzy kategoriami zmiennej porządkowej, to założenie monotoniczności wydaje się naturalne. Parametry niespełniające warunku 2.51 są trudne do interpretacji. Niemonotoniczna estymacja parametrów może być przypadkowa i wynikać z błędów losowych.

Podobnie jak w odniesieniu do modelu wierszowego uwzględnienie monotoniczności polega na ogół na przyjęciu założenia o równości parametrów, które zakłócają monotoniczność (Goodman 1985). Jeśli np.

$$u_1 < u_2 > u_3 < \quad \ldots < u_i \quad < \ldots < u_r \text{ oraz} \tag{2.52}$$

$$v_1 < v_2 < v_3 < \quad \ldots < v_j \quad < \ldots < v_c. \tag{2.53}$$

Można przyjąć, że $u_2 = u_3$. W odniesieniu do omawianego modelu sytuacja jest jednak bardziej skomplikowana. Przedstawiony wcześniej opis estymacji parametrów modelu za pomocą metod iteracyjnych pokazywał, że wielkości parametrów $v_i$ zależą od tego jak zostały oszacowane wielkości $u_i$, mówiąc inaczej parametry $v_i$ oraz $u_i$ są szacowane względem siebie. W związku z tym, jeśli przyjmiemy, że $u_2 = u_3$, może to wpłynąć na zakłócenie monotoniczności parametrów $v_i$. Wówczas możliwe jest wprowadzenie założenia o równości parametrów $v_i$, które zakłócają uporządkowanie, niemniej

strategia taka również nie gwarantuje rozwiązania problemu. Optymalną strategię, nakładania warunków zakładającą równość wybranych parametrów prezentują Ritov i Gilula (1991), jej przedstawienie wykracza jednak poza ramy tej pracy.

Choć na ogół parametry $u_i$ oraz $v_j$ są szacowane w modelu, możliwe jest samodzielne przypisanie „ocen" do kategorii obydwu zmiennych przez badacza (Agresti, 1984). Hipoteza taka była prostsza, gdyż w modelu szacowalibyśmy mniej parametrów. Postępowanie takie jest jednak zawsze do pewnego stopnia arbitralne, jeśli mamy do czynienia ze zmienną porządkową. Może to być uzasadnione w pewnych sytuacjach. Można wyobrazić sobie, że hipoteza, dotycząca badanego zjawiska zakłada coś na temat odległości pomiędzy kolejnymi kategoriami. Ponadto, jeśli jedna ze zmiennych jest mierzona na skali interwałowej, wykorzystanie tej informacji wydaje się naturalne. Należy pamiętać, że przypisanie wartości redukuje liczbę szacowanych parametrów i w konsekwencji zwiększa liczbę stopni swobody modelu. Przypisanie ocen w pewien określony sposób prowadzi do sformułowania modelu jednakowej interakcji bądź modelu efektu wierszowego (kolumnowego). Jak zostało zaznaczone powyżej, gdybyśmy określili, że odległości pomiędzy kolejnymi kategoriami są stałe w odniesieniu do obydwu zmiennych tj. $u_{i+1} - u_i = u$ oraz $v_{j+1} - v_j = v$ to omawiany model logarytmiczno–multiplikatywny byłby tożsamy z hipoteza o jednakowej interakcji, gdyż wszystkie lokalne stosunki szans byłyby sobie równe, tj.

$$\Theta_{ij}^{XY} = \delta^{uv}. \tag{2.54}$$

Gdybyśmy natomiast założyli jedynie, że odległości pomiędzy kategoriami zmiennej kolumnowej są takie same prowadziłoby to sformułowania hipotezy efektu wierszowego, gdyż:

$$\Theta_{ij}^{XY} = \delta^{(u_{i+1}-u_i)v} = \delta_i^v. \tag{2.55}$$

Podobnie nałożenie analogicznego warunku na parametry zmiennej wierszowej prowadzi do sformułowania efektu kolumnowego.

Model wierszowo–kolumnowy typu II - w odróżnieniu od prezentowanego wcześniej typu I - może być stosowany również do analizy związku pomiędzy zmiennymi nominalnymi. Wynika to z tego, że uporządkowanie kategorii jednej bądź drugiej zmiennej nie jest konieczne, gdyż parametry $u_i$ oraz $v_j$ są szacowane w modelu. O ile jednak w odniesieniu do zmiennych porządkowych jednowymiarowe skalowanie „ocen" wydaje się uzasadnione, to w przypadku zmiennych nominalnych - może być ono niewystarczające. Na przykład w odniesieniu do przynależności społeczno–zawodowej, rozstrzygnięcie, która kategoria zawodowa *rolnicy* czy *robotnicy wykwalifikowani* są grupą bliższą *inteligencji* zależy od tego jakie kryteria przyjmiemy. Goodman (1985, 1986) zaproponował rozszerzenie modelu wierszowo-kolumnowego dopusz-

czającego skalowanie $m-$wymiarowe zmiennych, tj. przypisanie kategoriom zmiennej parametrów $u_{i(m)}$ gdzie $m = 1, 2, ..M$. Jeśli przyjmiemy, że $M = 1$ wówczas mamy do czynienia ze zwykłym modelem wierszowo–kolumnowym omawianym poniżej. Maksymalna liczba wymiarów jest równa $M = r-1$ gdzie $r$ jest liczbą kategorii zmiennej. W przypadku uwzględnienia maksymalnej liczby wymiarów, mamy do czynienia z modelem nasyconym. Im mniej wymiarów uwzględniamy, tym mniej parametrów należy oszacować, a w konsekwencji liczba stopni swobody zdefiniowanego w ten sposób modelu jest większa.

## 2.4 Modele dla trzech zmiennych

Powyższa prezentacja pokazuje, że uwzględnienie porządkowego charakteru analizowanych zmiennych pozwala formułować dodatkowe hipotezy dotyczące związku pomiędzy dwiema zmiennymi. Podobnie się dzieje, gdy przedmiotem analizy jest rozkład większej liczby zmiennych. Ta część poświęcona zostanie modelom formułowanym dla trzech zmiennych $X$, $Y$, $Z$, w sytuacji, gdy jedna, dwie bądź trzy mierzone są na skali porządkowej. Modele te opisują związek pomiędzy zmiennymi odwołując się do podobnych hipotez, jak te, które zostały przedstawione powyżej: jednakowej interakcji, efektu wierszowego, itd.

W przypadku dwóch zmiennych uwzględnienie porządkowego charakteru jednej bądź dwóch zmiennych dotyczyło modelowania zależności pomiędzy zmiennymi, tj. stosunku szans $\Theta_{ij}^{XY}$ bądź — wyrażając to w języku parametrów modelu - interakcji $d_{ij}^{XY}$ pomiędzy zmiennymi. Nie inaczej będzie w przypadku rozkładu trzech zmiennych. W tym miejscu można raz jeszcze przywołać model nasycony dla trzech zmiennych:

$$\pi_{ijk}^{XYZ} = d \cdot d_i^X \cdot d_j^Y \cdot d_k^Z \cdot d_{ij}^{XY} \cdot d_{ik}^{XZ} \cdot d_{jk}^{YZ} \cdot d_{ijk}^{XYZ}. \tag{2.56}$$

Związek pomiędzy zmiennymi opisują parametry interakcji drugiego rzędu $d_{ij}^{XY}$, $d_{ik}^{XZ}$, $d_{jk}^{YZ}$, jak również parametr interakcji trzeciego rzędu $d_{ijk}^{XYZ}$. Przypomnijmy, że w przypadku zaprezentowanych w rozdziale pierwszym modeli dla zmiennych nominalnych, hipotezy dotyczyły jedynie tego czy związek istnieje, czy też nie, innymi słowy czy konieczne jest uwzględnienie interakcji dla danej pary zmiennych bądź interakcji trzeciego rzędu. W przypadku zmiennych porządkowych hipotezy dotyczą również tego, jakiego rodzaju jest to związek. O ile w przypadku modeli dla trzech zmiennych nominalnych możliwe było przedstawienie każdego modelu i omówienie jego własności, przedsięwzięcie takie jest praktycznie niewykonalne w odniesieniu do zmiennych porządkowych, gdyż można sformułować bardzo wiele takich modeli. Zostanie jedynie zasygnalizowane, jakiego typu modele możliwe są do sformułowania a szczegółowo

omówione zostaną wybrane z nich. Najpierw przedstawione zostaną modele uwzględniające jedynie interakcję drugiego rzędu. Następnie przedstawiona zostanie kwestia modelowanie interakcji trzeciego rzędu, czyli modele opisujące, w jaki sposób związek pomiędzy dwiema zmiennymi może zależeć od wartości trzeciej zmiennej.

### 2.4.1 Modelowanie interakcji drugiego rzędu

W tej części skupię się na hipotezach, które zakładają, że związek pomiędzy dwiema zmiennymi nie zależy od wartości, jakie przyjmuje trzecia zmienna. Wyrażając to za pomocą równości warunkowych stosunków szans można przywołać warunek 1.28, jak również zapisać go w nieco innej postaci:

$$\begin{aligned} \Theta_{ij(k)}^{XY(Z)} &= \delta_{ij}^{XY}, \\ \Theta_{i(j)k}^{X(Y)Z} &= \delta_{ik}^{XZ}, \\ \Theta_{(i)jk}^{(X)YZ} &= \delta_{jk}^{YZ}. \end{aligned} \tag{2.57}$$

Jak łatwo zauważyć powyższe warunki definiują analizowany w pierwszym rozdziale model $[XY][XZ][YZ]$. Wszystkie stosunki szans są specyficzne dla kombinacji kategorii dwóch zmiennych, których dotyczą, ale ich wielkość nie zależy od wartości trzeciej zmiennej, która określa podzbiorowość, dla której stosunek szans został wyznaczony (w powyższych wzorach ta zmienna podana jest w nawiasie). Modele przedstawione poniżej będą szczególnym przypadkiem tej sytuacji. Skupimy się na modelach hierarchicznych, tak więc aby było możliwe uwzględnienie interakcji dwóch zmiennych (np. $d_{ij}^{XY}$), należy uwzględnić w modelu efekty tych zmiennych (odpowiednio $d_i^X$ oraz $d_j^Y$). Nie będą więc rozważane modele prostsze niż $[XY]$, $[XZ]$ oraz $[YZ]$.

Zanim omówione zostaną ogólne możliwości formułowania hipotez dotyczących interakcji drugiego rzędu zostanie podany przykład stosunkowo prostej hipotezy. Przyjmijmy, że zmienne $X$ oraz $Y$ są mierzone na skali porządkowej. Hipoteza głosi, że:

$$\begin{aligned} \Theta_{ij(k)}^{XY(Z)} &= \delta, \\ \Theta_{i(j)k}^{X(Y)Z} &= 1, \\ \Theta_{(i)jk}^{(X)YZ} &= 1, \end{aligned} \tag{2.58}$$

dla dowolnych $i$, $j$, $k$. Zgodnie z tą hipotezą:

1. Zależność pomiędzy zmiennymi $X$ i $Y$ w każdej podzbiorowości wyróżnionej ze względu na wartości trzeciej zmiennej $Z$ daje się opisać za pomocą pojęcia jednakowej interakcji. Co więcej, siła tej zależności jest taka sama dla każdej wartości $Z$ .

2. Zmienne $X$ i $Z$ są warunkowo niezależne stochastycznie względem $Y$.

3. Zmienne $Y$ i $Z$ są warunkowo niezależne stochastycznie względem $X$.

Jest to więc szczególny przypadek modelu $[XY][Z]$. O ile w modelu $[XY][Z]$ wzór zależności pomiędzy $X$ i $Y$ jest nieokreślony i wymaga $(r-1)(c-1)$ parametrów, wprowadzony model opisuje tę zależność w sposób prostszy, za pomocą jednego parametru. Oznaczmy ten model jako $[XY_{UA}][Z]$. Z drugiej strony jest on nieznacznie bardziej złożony niż model niezależności trzech zmiennych: głosi, że wszystkie stosunki szans między zmiennymi $X$ oraz $Y$ są sobie równe, natomiast niekoniecznie równe 1. W stosunku do tego modelu liczba stopni swobody jest mniejsza o 1 i wynosi $df = rct - r - c - t + 1$. Ilustracją tej hipotezy jest tabela 2.17, która podobnie jak tabele 2.4, 2.8, 2.12, 2.15 uwzględnia jedynie parametry interakcji. Jak widać w każdej podzbiorowości wyróżnionej ze względu na zmienną $Z$ wszystkie lokalne stosunki szans są sobie równe.

Tabela 2.17: Ilustracja modelu $[XY_{UA}][Z]$ - parametry interakcji

| | $Z=z_1$ | | | | $Z=z_2$ | | | | $Z=z_3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
| $x_1$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\delta$ | $\delta^2$ | $\delta^3$ | 1 | $\delta$ | $\delta^2$ | $\delta^3$ | 1 | $\delta$ | $\delta^2$ | $\delta^3$ |
| $x_3$ | 1 | $\delta^2$ | $\delta^4$ | $\delta^6$ | 1 | $\delta^2$ | $\delta^4$ | $\delta^6$ | 1 | $\delta^2$ | $\delta^4$ | $\delta^6$ |
| $x_4$ | 1 | $\delta^3$ | $\delta^6$ | $\delta^9$ | 1 | $\delta^3$ | $\delta^6$ | $\delta^9$ | 1 | $\delta^3$ | $\delta^6$ | $\delta^9$ |

Tabela 2.17 jest zgodna z parametryzacją względem kategorii odniesienia $x_1$, $y_1$ przy dowolnej kategorii odniesienia zmiennej $Z$. Ogólnie model ten można przedstawić jako:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \delta^{(i-a)(j-b)}, \tag{2.59}$$

gdzie kategoriami odniesienia są $x_a$, $y_b$. Powyższą hipotezę można również uprościć, jeśli założy się dodatkowo, że rozkład zmiennej $Z$ jest równomierny w pozbiorowościach wyróżnionych ze względu na zmienne $Y$ oraz $X$. Taki model oznaczmy jako $[XY_{UA}]$ a formuła na prawdopodobieństwa oczekiwane zgodne z tym modelem przedstawiałaby się następująco:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \delta^{(i-a)(j-b)}. \tag{2.60}$$

Liczba stopni swobody dla tego modelu wynosi $df = rc(t-1) - 1$, tj. jest o jeden mniejsza niż w modelu $[XY]$.

Jak widać logika formułowania hipotez dla trzech zmiennych porządkowych jest podobna jak w przypadku zmiennych nominalnych. Niemniej, dzięki uwzględnieniu uporządkowania kategorii zmiennej (bądź zmiennych) interakcję drugiego rzędu można modelować na wiele sposobów, tj. można wyróżnić wiele modeli będących szczególnym przypadkiem hipotezy $[XY][XZ][YZ]$. Interakcję dla dowolnej pary zmiennych można modelować w za pomocą hipotezy o:

1. Niezależności stochastycznej (NULL),
2. Jednakowej interakcji (UA),
3. Efekcie wierszowym (R),
4. Efekcie kolumnowym (C),
5. Logarytmiczno–liniowym efekcie wierszowo–kolumnowym (RC1),
6. Logarytmiczno–multiplikatywnym efekcie wierszowo–kolumnowym (RC2),
7. Nieokreślonym wzorze zależności (FA-*full association*) - brak hipotezy.

Tabela 2.18: Rózne typy interakcji drugiego rzędu

| Para zmiennych | $X$ i $Y$ | $X$ i $Z$ | $Y$ i $Z$ |
|---|---|---|---|
| Typ interakcji | $\Theta_{ij(k)}^{XY(Z)} = \ldots$ | $\Theta_{i(j)k}^{X(Y)Z} = \ldots$ | $\Theta_{(i)jk}^{(X)YZ} = \ldots$ |
| NULL | 1 | 1 | 1 |
| UA | $\delta^{XY}$ | $\delta^{XZ}$ | $\delta^{YZ}$ |
| R | $\delta_{i\cdot}^{XY}$ | $\delta_{i\cdot}^{XZ}$ | $\delta_{j\cdot}^{YZ}$ |
| C | $\delta_{\cdot j}^{XY}$ | $\delta_{\cdot k}^{XZ}$ | $\delta_{\cdot k}^{YZ}$ |
| RC1 | $\delta_{i\cdot}^{XY}\delta_{\cdot j}^{XY}$ | $\delta_{i\cdot}^{XZ}\delta_{\cdot k}^{XZ}$ | $\delta_{j\cdot}^{YZ}\delta_{\cdot k}^{YZ}$ |
| RC2 | $\left(\delta^{XY}\right)^{(u_{i+1}-u_i)(v_{j+1}-v_j)}$ | $\left(\delta^{XZ}\right)^{(u_{i+1}-u_i)(w_{k+1}-w_k)}$ | $\left(\delta^{YZ}\right)^{(v_{j+1}-v_j)(w_{k+1}-w_k)}$ |
| FA | $\delta_{ij}^{XY}$ | $\delta_{ik}^{XZ}$ | $\delta_{jk}^{YZ}$ |

Tabela 2.18 prezentuje, w jaki sposób można modelować interakcję pomiędzy każdą parą zmiennych. Tabela ta pokazuje, ile wynoszą warunkowe stosunki szans przy każdym typie interakcji. Przypomnijmy, że w przypadku, gdy wszystkie, trzy zmienne są nominalne, każda z interakcji może być modelowana tylko na trzy sposoby: można założyć, że interakcja nie występuje (NULL), interakcja ma charakter nieokreślony (FA), ewentualnie można zastosować logarytmiczno–multiplikatywny model

wierszowo–kolumnowy (RC2), który nie wymaga uporządkowania kategorii zmiennej. W przypadku uwzględnienia porządkowego charakteru zmiennych interakcję dla każdej pary zmiennych daje się modelować na siedem sposobów, tak więc możliwości formułowania hipotez możliwości są znacznie większe. Aby zdać sobie sprawę, jakie hipotezy są możliwe do sformułowania, konieczne jest poczynienie kilku dodatkowych uwag.

Istotne jest, ile spośród analizowanych zmiennych mierzonych jest na skali porządkowej: jedna, dwie czy też wszystkie trzy zmienne. W przypadku, gdy zmienną tą jest wyłącznie jedna zmienna (np. $X$) można dla niej formułować modele wierszowe (bądź kolumnowe) z pozostałymi dwiema zmiennymi ($Y$ i $Z$). Nie wchodzą jednak w grę modele jednakowej interakcji. Ponadto, dla pozostałych dwóch zmiennych ($Y$ i $Z$) nie jest możliwe formułowanie żadnych „dodatkowych" hipotez.

Warto również zauważyć, że przyjęcie pewnych dodatkowych założeń dla dwóch różnych par interakcji prowadzi do sformułowania kolejnych hipotez. Rozważmy na przykład hipotezę, która głosi, że interakcja pomiędzy parami zmiennych $X$ oraz $Z$ jak również $Y$ oraz $Z$ jest opisywana za pomocą jednakowej interakcji, natomiast nie zakładamy nic na temat związku pomiędzy $X$ i $Y$. Hipotezę tę można zapisać jako:

$$\begin{aligned} \Theta_{ij(k)}^{XY(Z)} &= \delta_{ij}^{XY}, \\ \Theta_{i(j)k}^{X(Y)Z} &= \delta^{XZ}, \\ \Theta_{(i)jk}^{(X)YZ} &= \delta^{YZ}. \end{aligned} \tag{2.61}$$

Model ten oznaczamy jako zapisać jako $[XY][XZ_{UA}][YZ_{UA}]$ i zgodnie z parametryzacją względem kategorii $x_a$, $y_b$, $y_c$ odniesienia można zapisać go jako:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ij}^{XY} \left(\delta^{XZ}\right)^{(i-a)(k-c)} \cdot \left(\delta^{YZ}\right)^{(j-b)(k-c)}. \tag{2.62}$$

Model $[XY][XZ_{UA}][YZ_{UA}]$ posiada w stosunku do modelu $[XY][XZ][YZ]$ dodatkowe założenia dotyczące równości wszystkich lokalnych stosunków szans, jakie można wyznaczyć dla zmiennych $X$ oraz $Z$, jak również identyczności lokalnych stosunków szans dla zmiennych $Y$ oraz $Z$. Tak więc liczba stopni swobody dla tego modelu wynosi

$$df = (r-1)(c-1)(t-1) + (r-1)(t-1) - 1 + (c-1)(t-1) - 1 = (t-1)(rc-1) - 2.$$

W tabeli 2.19 przedstawiamy rozkład trzech zmiennych: opinie dotyczące tego, czy rząd powinien zmniejszyć różnice w dochodach ($X$) i tego czy rząd powinien zapewnić każdemu pracę ($Y$) oraz oceny sytuacji materialnej własnego gospodarstwa domowego ($Z$)[11].

[11]Dokładne brzmienie pytań dotyczących zmiennych $X$ oraz $Z$ zostało podane wcześniej przy okazji analizowania danych z tabeli 2.5. Jeśli chodzi o zmienną $Y$ respondenci musieli ustosunkować

Tabela 2.19: Zadowolenie z własnej sytuacji materialnej a opinie dotyczące działań rządu[a]

| Osoby „zadowolone" z własnej sytuacji finansowej ($Z = 1$) | | | |
|---|---|---|---|
| Rząd powinien zmniejszyć różnice w dochodach ($X$) | Rząd powinien zagwarantować pracę ($Y$) | | |
| | 1. Zgadzam się | 2. Ani się zgadzam, ani nie zgadzam | 3. Nie zgadzam się |
| 1. Zgadzam się | 93,0 | 9,9 | 4,1 |
| 2. Ani się zgadzam ani nie zgadzam | 8,4 | 2,9 | 2,9 |
| 3. Nie zgadzam się | 7,1 | 2,6 | 8,2 |
| Osoby „mniej więcej zadowolone" z własnej sytuacji finansowej ($Z = 2$) | | | |
| 1. Zgadzam się | 328,6 | 14,2 | 13,1 |
| 2. Ani się zgadzam ani nie zgadzam | 26,8 | 8,7 | 7,0 |
| 3. Nie zgadzam się | 12,7 | 4,7 | 17,4 |
| Osoby „niezadowolone" z własnej sytuacji finansowej ($Z = 3$) | | | |
| 1. Zgadzam się | 575,7 | 15,2 | 8,4 |
| 2. Ani się zgadzam ani nie zgadzam | 30,4 | 5,0 | 1,1 |
| 3. Nie zgadzam się | 13,4 | 1,0 | 10,0 |

[a]Źródło: Polski Generalny Sondaż Społeczny, 2005. Dane przeważone.

Okazuje się, że model $[XY][XZ_{UA}][YZ_{UA}]$ jest akceptowalny na poziomie istotności 0,05: $G^2 = 13,33$ przy 14 stopniach swobody ($p = 0,50$), $\chi^2 = 13,08$ ($p = 0,52$). Wartości parametrów jednakowej interakcji wynoszą odpowiednio: $\delta^{XZ} = 0,72$, $\delta^{YZ} = 0,80$.

W odniesieniu do powyższych danych możliwe jest skonstruowanie hipotezy prostszej: można założyć, że siła interakcji pomiędzy zmienną $X$ a zmienną $Z$ (mierzona wielkością lokalnego stosunku szans) jest taka sama jak interakcja pomiędzy $Y$ i $Z$. In-

się do twierdzenia: *Rząd powinien zapewnić każdemu pracę, kto chce pracować* na pięciopunktowej skali *1. Zdecydowanie się zgadzam, 2. Zgadzam się, 3. Ani się zgadzam, ani nie zgadzam, 4. Nie zgadzam się, 5. Zdecydowanie się nie zgadzam*. Ze względu na konieczność zapewnienia odpowiednich liczebności komórkom rozkładu łącznego, połączone zostały kategoria 1 i 2 oraz kategoria 4 i 5 (dotyczyło to również zmiennej $X$ mierzonej na takiej samej skali). Osoby, które nie udzieliły odpowiedzi na przynajmniej jedno z pytań zostały wyłączone z analizy (stanowiły one 2,4% wszystkich respondentów).

nymi słowy w stosunku do poprzedniej hipotezy zakładamy dodatkowo, że $\delta^{XZ} = \delta^{YZ}$. Model taki możemy oznaczyć jako $[XZ_{UA}]$=$[YZ_{UA}]$ i przedstawić następująco:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ij}^{XY} \cdot \delta^{(i-a)(k-c)} \cdot \delta^{(j-b)(k-c)}. \tag{2.63}$$

Liczba niezależnych parametrów dla tego modelu jest mniejsza o jeden niż w modelu poprzednim, co wynika z dodatkowego założenia $\delta^{XZ} = \delta^{YZ}$. Liczba stopni swobody jest więc o 1 większa. Dla danych z tabeli 2.19 statystyka $G^2 = 13,75$ jest nieznacznie większa niż w modelu poprzednim, $df = 15$ ($p = 0,54$). Parametr jednakowej interakcji $\delta$ jest — zgodnie z tym modelem — taki sam dla par zmiennych $X$ i $Z$ oraz $Y$ i $Z$ i wynosi 0,76. Test warunkowy pokazuje, że założenie o równości tych parametrów nie pogarsza dopasowania w sposób istotny statystycznie: $G^2 = 0,22$ przy jednym stopniu swobody ($p = 0,64$). Wartość parametru jednakowej interakcji dla tego modelu można interpretować następująco: proporcja osób, które zgadzają się z tym, że rząd powinien zmniejszyć różnice w dochodach do takich, które są neutralne wobec takiej opinii jest 1,32 (1/0,76) razy mniejsza, wśród osób, które są *zadowolone* z sytuacji finansowej aniżeli wśród *średnio zadowolonych*. Wniosek taki możemy oczywiście wypowiedzieć dla dowolnej pary „sąsiednich" kategorii zmiennej $X$ i zmiennej $Z$ w każdej podzbiorowości wyróżnionej ze względu na zmienną $Y$. Warunek $\delta^{XZ} = \delta^{YZ}$, który dodatkowo nałożyliśmy pozwala sformułować podobny wniosek odnośnie analogicznych kategorii zmiennej opisującej opinię dotyczącego, tego czy rząd powinien zapewnić wszystkim pracę.

Powyższy przykład pokazuje, że jest możliwe formułowanie prostszych modeli poprzez nakładanie pewnych dodatkowych warunków, tj. przyrównanie do siebie parametrów interakcji dla jednej pary zmiennych do parametrów interakcji dla innej pary zmiennej Należy jednak zwrócić uwagę, że w odniesieniu do danych z tabeli (2.19) było to sensowne, gdyż pytania dotyczące opinii w sprawie zagwarantowania pracy i zmniejszenia różnic w dochodach przez rząd skonstruowane były w podobny sposób: respondenci posługiwali się tą samą skalą odpowiedzi[12].

W odniesieniu do modelu $[XY][XZ_{UA}][YZ_{UA}]$ założenie o równości odpowiednich parametrów jest możliwe, choć nie jest konieczne. Wskażemy teraz sytuację, w której poczynienie takiego założenia jest szczególnie istotne. Przypuśćmy, że związek $X$ i $Z$ oraz $Y$ i $Z$ modelujemy za pomocą logarytmiczno–multiplikatywnej interakcji wierszowo–kolumnowej (RC2). Oznacza to, że do modelu wprowadzamy parametry

[12]Formułowanie tego rodzaju hipotezy jest możliwe jest również w odniesieniu do trzech zmiennych nominalnych. W modelu $[XZ][YZ]$, interakcja pomiędzy $X$ i $Z$ oraz $Y$ i $Z$ ma nieokreślony charakter można jednak założyć, że analogiczne stosunki szans są sobie równe tj. $\Theta_{ik}^{XZ} = \Theta_{ik}^{YZ}$.

skalujące wartości poszczególnych zmiennych. Hipotezę taką można zapisać jako:

$$\begin{aligned} \Theta^{XY(Z)}_{ij(k)} &= \delta^{XY}_{ij}, \\ \Theta^{X(Y)Z}_{i(j)k} &= \left(\delta^{XZ}\right)^{(u_{i+1}-u_i)(w_{k+1}-w_k)}, \\ \Theta^{(X)YZ}_{(i)jk} &= \left(\delta^{YZ}\right)^{(v_{j+1}-v_j)(t_{k+1}-t_k)}. \end{aligned} \tag{2.64}$$

Zauważmy, że parametry związane ze zmienną $Z$ można skalować dwukrotnie: ze względu na interakcję ze zmienną $X$ (parametry $w_k$) i ze względu na interakcję ze względu na zmienną $Y$, tj: parametry (parametry $t_k$). Interpretacja takiej hipotezy byłaby trudna. Wydaje się, w takim przypadku sensowne jest przyjęcie założenia: $w_k = t_k$. Model taki jest bardzo dobrze dopasowany do analizowanych danych, odpowiednie statystyki wynoszą $\chi^2 = 6,21$ $G^2 = 6,48$ przy 11 stopniach swobody[13]. (wielkości $p$ wynoszą odpowiednio 0,86 i 0,84). Parametry tego modelu wynoszą odpowiednio $\delta^{XZ} = 0,43$, $\delta^{YZ} = 0,67$ a parametry skalujące wynoszą, $u_1 = -0,79$, $u_2 = 0,22$, $u_3 = 0,56$ dla zmiennej $X$, oraz $v_1 = -0,82$, $v_2 = 0,41$, $v_3 = 0,41$ dla kategorii zmiennej $Y$[14]; $w_1 = 0,60$, $w_2 = 0,17$, $w_3 = -0,78$ dla zmiennej $Z$. Parametry tego modelu sugerują, że odległość pomiędzy kategorią *zadowolony* ze swoich dochodów i *mniej więcej zadowolony* jest w przybliżeniu dwukrotnie mniejsza niż odległość pomiędzy kategoriami *mniej więcej zadowolony* i *niezadowolony* ($w_1 - w_2 = 0,43$, natomiast $w_2 - w_3 = 0,95$).

Spójność skalowania zmiennej w interakcjach z różnymi zmiennymi — tak jak w powyższym przykładzie — może też odnosić się do łączenia logarytmiczno–multiplikatywnej interakcji wierszowo–kolumnowej (RC2) z jednakową interakcją (UA). Przypomnijmy, że o ile w modelu wierszowo–kolumnowym skalowanie obydwu zmiennych daje takie same odległości pomiędzy kolejnymi kategoriami zmiennej, tj. $u_i - u_{i+1} = const$ model ten jest tożsamy z modelem jednakowej interakcji. Dlatego, jeśli zakładamy, jednakową interakcję pomiędzy zmiennymi $X$ i $Z$, to jeśli chcemy założyć, że interakcja pomiędzy zmienną $Y$ i $Z$ jest opisywana za pomocą interakcji wierszowo-kolumnowej sensownie jest założyć, że odległości pomiędzy kolejnymi kategoriami zmiennej $Z$ są takie same.

Ze względu na wielość modeli, jakie można sformułować wykorzystując ich porządkowy charakter nie wydaje się celowe wymienienie ich wszystkich i analizowanie własności każdego z nich. Powyżej omówione zostały w sposób ogólny możliwości ich

[13]Poza parametrami interakcji w modelu występuje $(r-2)$ niezależnych parametrów związanych ze zmienną $X$, $(c-2)$ związanych ze zmienną $Y$ i $(t-2)$ związanych ze zmienną $Z$.

[14]Pierwotnie, parametry dla kategorii zmiennej $Y$, która jest zmienną porządkową były niemonotoniczne, choć zakłócenie monotoniczności było nieznaczne: $v_1 = -0,81$, $v_2 = 0,41$, $v_3 = 0,40$. Wprowadzenie dodatkowego założenia $v_2 = v_3$ (bardziej ogólnie $v_2 \leq v_3$) praktycznie nie zmieniło wielkości pozostałych parametrów modelu i na wpłynęło na jego dopasowanie do danych.

Tabela 2.20: Wykształcenie, ocena własnej sytuacji materialnej i opinie dotyczące zmniejszenia zróżnicowania w dochodach[a]

| Osoby *zadowolone* z sytuacji finansowej ($Z = 1$) | | | |
|---|---|---|---|
| | Rząd powinien zmniejszyć różnice w dochodach ($Y$) | | |
| Wykształcenie ($X$) : | 1. Zgadzam się | 2. Ani się zgadzam, ani nie zgadzam | 3. Nie zgadzam się |
| 1. Podstawowe | 40,6 | 4,6 | 5,7 |
| 2. Niepełne średnie | 32,6 | 12,1 | 8,2 |
| 3. Ukończone średnie | 32,8 | 9,9 | 10,4 |
| 4. Wyższe | 21,1 | 7,8 | 21,7 |
| Osoby *mniej więcej zadowolone* z sytuacji finansowej ($Z = 2$) | | | |
| 1. Podstawowe | 166,2 | 15,0 | 11,6 |
| 2. Niepełne średnie | 156,7 | 29,8 | 10,6 |
| 3. Ukończone średnie | 149,0 | 37,7 | 26,5 |
| 4. Wyższe | 69,0 | 27,4 | 39,8 |
| Osoby *niezadowolone* z sytuacji finansowej ($Z = 3$) | | | |
| 1. Podstawowe | 317,4 | 25,5 | 9,4 |
| 2. Niepełne średnie | 285,5 | 23,8 | 16,4 |
| 3. Ukończone średnie | 179,6 | 26,2 | 17,5 |
| 4. Wyższe | 65,3 | 23,2 | 14,9 |

[a]Źródło: Polski Generalny Sondaż Społeczny, 2007. Dane przeważone.

formułowania: jakiego typu hipotezy można formułować dla interakcji pomiędzy daną parą zmiennych, czy hipoteza odnośnie jednej pary zmiennych ma konsekwencje dla formułowania hipotezy dla innej pary zmiennych. Dopełnieniem będzie przykład empiryczny. W tabeli 2.20 przedstawiony został rozkład łączny trzech zmiennych: wykształcenia ($X$), tego czy rząd powinien zmniejszyć różnice w dochodach ($Y$) i oceny sytuacji materialnej własnego gospodarstwa domowego ($Z$)[15] (PGSS, 1997).

[15]Konstrukcja pytań o sytuację materialną ($Z$) i opinii dotyczącej zmniejszenia różnicy w dochodach ($Y$) jest identyczna jak w przedstawionych wcześniej analizach (porównaj informacje podane przy okazji omówienia tabel 2.5 oraz 2.19). W pytaniu o wykształcenie ($X$) wyróżniono 10 kategorii: 0–brak wykształcenia szkolnego, 1–niepełne podstawowe, 2–podstawowe, 3–zasadnicze zawodowe, 4–niepełne średnie, 5–średnie ogólnokształcące, 6–średnie zawodowe, 7–pomaturalne, 8–nieukończone wyższe, 9–ukończone wyższe. Ze względu na małe liczebności kategorie połączyliśmy w następujący sposób: 1–2, 3–4, 5–6, 7–9. W analizach zakładamy, że utworzone w ten sposób cztery kategorie wykształcenia tworzą pewną hierarchię, tak więc mamy do czynienia ze skalą porządkową. Z analizy wykluczone zostały osoby, które w momencie miały mniej niż 25 lat, gdyż można oczekiwać, że

Analizując te dane porównywane będą ze sobą konkurencyjne hipotezy jak również interpretacja wybranych modeli.

Tabela 2.21: Wyniki weryfikacji hipotez dla danych z tabeli 2.20

| Model | df | $\chi^2$ | $G^2$ | $\Delta$ |
|---|---|---|---|---|
| $[XY]$ | 24 | 622,9 ($p < 0,0001$) | 706,9 ($p < 0,0001$) | 23,8 |
| $[XZ]$ | 24 | 1889,6 ($p < 0,0001$) | 1825,1 ($p < 0,0001$) | 44,6 |
| $[YZ]$ | 27 | 290,0 ($p < 0,0001$) | 325,3 ($p < 0,0001$) | 16,0 |
| $[X][Y][Z]$ | 28 | 360,0 ($p < 0,0001$) | 267,8 ($p < 0,0001$) | 12,8 |
| $[XY][Z]$ | 22 | 122,6 ($p < 0,0001$) | 118,3 ($p < 0,0001$) | 9,7 |
| $[XZ][Y]$ | 22 | 240,2 ($p < 0,0001$) | 207,9 ($p < 0,0001$) | 10,9 |
| $[YZ][X]$ | 24 | 211,7 ($p < 0,0001$) | 194,1 ($p < 0,0001$) | 10,5 |
| $[XY][XZ]$ | 16 | 62,6 ($p < 0,0001$) | 58,5 ($p < 0,0001$) | 5,4 |
| $[XY][YZ]$ | 18 | 45,1 ($p = 0,0004$) | 44,7 ($p = 0,0005$) | 6,2 |
| $[XZ][YZ]$ | 18 | 146,1 ($p < 0,0001$) | 134,3 ($p < 0,0001$) | 8,3 |
| $[XY][XZ][YZ]$ | 12 | 11,4 ($p = 0,4971$) | 11,4 ($p = 0,4978$) | 19,3 |
| $[XY_{UA}][XZ_{UA}][YZ_{UA}]$ | 25 | 28,7 ($p = 0,2775$) | 28,5 ($p = 0,2864$) | 4,2 |
| $[XY_{UA}][YZ_{UA}]$ | 26 | 85,4 ($p < 0,0001$) | 76,1 ($p < 0,0001$) | 6,5 |
| $[XY_{RC2}][XZ_{RC2}][YZ_{RC2}]^a$ | 21 | 18,7 ($p = 0,6047$) | 18,7 ($p = 0,6068$) | 3,0 |
| $[XY_{RC1}][XZ_{RC1}][YZ_{RC1}]$ | 17 | 14,0 ($p = 0,6676$) | 14,0 ($p = 0,6665$) | 2,4 |
| $[XY_{RC1}][XZ_{UA}][YZ_{UA}]$ | 22 | 21,8 ($p = 0,4691$) | 21,9 ($p = 0,4661$) | 3,7 |
| $[XY_{UA}][XZ_{RC1}][YZ_{UA}]$ | 22 | 20,0 ($p = 0,5843$) | 19,8 ($p = 0,5928$) | 2,9 |
| $[XY_{UA}][XZ_{UA}][YZ_{RC1}]$ | 23 | 28,3 ($p = 0,2048$) | 28,1 ($p = 0,2119$) | 4,2 |
| $[XY_R][XZ_{UA}][YZ_{UA}]$ | 23 | 23,9 ($p = 0,4110$) | 23,8 ($p = 0,4139$) | 3,9 |
| $[XY_C][XZ_{UA}][YZ_{UA}]$ | 24 | 27,4 ($p = 0,2876$) | 27,4 ($p = 0,2838$) | 4,3 |
| $[XY_{UA}][XZ_R][YZ_{UA}]$ | 23 | 26,1 ($p = 0,2971$) | 25,9 ($p = 0,3047$) | 3,8 |
| $[XY_{UA}][XZ_C][YZ_{UA}]$ | 24 | 23,3 ($p = 0,5036$) | 23,2 ($p = 0,5056$) | 4,0 |
| $[XY_R][XZ_C][YZ_{UA}]$ | 22 | 18,4 ($p = 0,6809$) | 18,4 ($p = 0,6802$) | 3,4 |

[a]W modelu zakłada się identyczne skalowanie zmiennej w interakcjach z różnymi zmiennymi.

Tabela 2.21 pokazuje, że modele proste modele takie jak $[XY]$, $[YZ]$ oraz $[XZ]$, które nie uwzględniają efektów wszystkich trzech zmiennych nie są akceptowalne na

osoby takie nie zakończyły jeszcze procesu kształcenia. Łącznie z analizy wykluczono prawie 19% respondentów ze zrealizowanej próby (uwzględniając również osoby, które na pytania związane ze zmiennymi $Z$ oraz $Y$ nie udzieliły odpowiedzi bądź odpowiedziały „Trudno powiedzieć").

poziomie istotności równym $\alpha = 0,05$ (dla ustalenia uwagi taki poziom istotności przyjęty został przy podejmowaniu decyzji o odrzuceniu bądź akceptacji kolejnych modeli i testach warunkowych). Trudno również oczekiwać, że wszystkie trzy zmienne są niezależne stochastycznie. Potwierdzają to dane: model $[X][Y][Z]$, nie może być zaakceptowany. Pozostaje pytanie o interakcję drugiego i trzeciego stopnia pomiędzy zmiennymi. Wyniki weryfikacji modelu $[XY][XZ][YZ]$ pokazują, ze model taki jest akceptowalny,tak więc nie jest konieczne uwzględnianie interakcji trzeciego rzędu. Powstaje pytanie: czy związku pomiędzy zmiennymi nie można opisać za pomocą modeli prostszych?

W pierwszej kolejności można zapytać, czy konieczne jest uwzględnienie interakcji drugiego rzędu dla każdej pary zmiennych. Można na przykład przypuszczać, że wykształcenie i opinie dotyczące zmniejszania zróżnicowania dochodów są niezależne stochastycznie w każdej podzbiorowości wyróżnionej ze względu na ocenę sytuacji materialnej. Hipoteza taka odpowiadałaby modelowi $[XZ][YZ]$. Okazuje się jednak, że model taki nie może być zaakceptowany, jeśli posługujemy się statystykami $G^2$ oraz $\chi^2$. Podobnie jest z hipotezami sformułowanymi analogiczne w odniesieniu do każdej innej pary zmiennych: należałoby odrzucić modele $[XY][XZ]$, $[XY][YZ]$, dla obydwu $p < 0,001$. Modele, w których dopuszcza się, że tylko jedna para zmiennych jest zależna stochastycznie, tj. $[XY][Z]$, $[XZ][Y]$, $[YZ][X]$, również nie mogą być zaakceptowane.

Zgodnie z nakreśloną powyżej strategią, okazuje się, że w modelu powinniśmy uwzględnić interakcję drugiego rzędu dla każdej pary zmiennych. Przedstawione powyżej hipotezy nie zakładały nic odnośnie typu zależności. Wykorzystując fakt, że wszystkie zmienne mierzone są skali porządkowej, warto postawić pytanie, czy możliwe jest opisanie tej zależności w sposób prostszy aniżeli w modelu $[XY][XZ][YZ]$. Można oczekiwać na przykład, że związek pomiędzy wykształceniem i opiniami dotyczącymi zmniejszania zróżnicowania dochodów w każdej pozbiorowości wyróżnionej ze względu na ocenę własnej sytuacji materialnej daje się opisać za pomocą hipotezy jednakowej interakcji. Podobne założenie można poczynić dla dwóch pozostałych par zmiennych. Hipotezę taką oznaczymy $[XY_{UA}][XZ_{UA}][YZ_{UA}]$. Zgodnie z nią:

$$\begin{aligned} \Theta^{XY(Z)}_{ij(k)} &= \delta^{XY}, \\ \Theta^{X(Y)Z}_{i(j)k} &= \delta^{XZ}, \\ \Theta^{(X)YZ}_{(i)jk} &= \delta^{YZ}. \end{aligned} \tag{2.65}$$

Zauważmy, że w stosunku do hipotezy o niezależności stochastycznej trzech zmiennych omawiany model posiada jedynie trzy parametry więcej. Jak pokazują wyniki

przedstawione w tabeli 2.21 model ten jest akceptowalny: przy 25 stopniach swobody $\chi^2 = 28,7$ $(p = 0,28)$ oraz $G^2 = 28,5$ $(p = 0,29)$.

Tabela 2.22: Rozkład oczekiwany dla danych z tabeli 2.20 zgodny z modelem zakładającym jednakową interakcję dla każdej pary zmiennych

| Osoby *zadowolone* z sytuacji materialnej ($Z = 1$) | | | |
|---|---|---|---|
| | Rząd powinien zmniejszyć różnice w dochodach ($Y$) | | |
| Wykształcenie ($X$): | 1. Zgadzam się | 2. Ani się zgadzam, ani nie zgadzam | 3. Nie zgadzam się |
| 1. Podstawowe | 31,9 | 5,1 | 3,1 |
| 2. Niepełne średnie | 36,8 | 8,6 | 7,7 |
| 3. Ukończone średnie | 34,9 | 11,9 | 15,8 |
| 4. Wyższe | 20,9 | 10,5 | 20,4 |
| Osoby *mniej więcej zadowolone* z sytuacji materialnej ($Z = 2$) | | | |
| 1. Podstawowe | 173,9 | 19,3 | 8,3 |
| 2. Niepełne średnie | 169,4 | 27,6 | 17,3 |
| 3. Ukończone średnie | 135,9 | 32,4 | 29,9 |
| 4. Wyższe | 68,7 | 24,1 | 32,6 |
| Osoby *niezadowolone* z sytuacji materialnej ($Z = 3$) | | | |
| 1. Podstawowe | 322,1 | 24,9 | 7,4 |
| 2. Niepełne średnie | 265,1 | 30,1 | 13,2 |
| 3. Ukończone średnie | 179,6 | 29,9 | 19,2 |
| 4. Wyższe | 76,7 | 18,7 | 17,7 |

Tabela 2.22 przedstawia rozkład oczekiwany zgodny z tą hipotezą. Dokładne przestudiowanie tej tabeli pokazuje, że wszystkie lokalne stosunki szans dla zmiennych $X$ i $Y$ są sobie równe, co więcej są one takie same dla każdej podzbiorowości wyróżnionej ze względu na zmienną $Z$. Tj.

$$\frac{31,9 \cdot 8,6}{5,1 \cdot 36,8} = \frac{5,1 \cdot 7,7}{3,1 \cdot 8,6} = \ldots = \frac{173,9 \cdot 27,6}{19,3 \cdot 169,4} = \ldots = \delta^{XY} = 1,46.$$

To samo można powiedzieć o zależności pomiędzy pozostałymi parami zmiennych, przy czym odpowiednie parametry jednakowej interakcji są równe: $\delta^{XZ} = 0,85$, $\delta^{YZ} = 0,69$. Okazuje się, że stosunkowo najsłabsza jest warunkowa zależność pomiędzy wykształceniem i zadowoleniem z sytuacji materialnej. Wartość $\delta^{XZ}$ jest stosunkowo bliska wartości 1, czyli takiej wartości parametru który opisuje stan warunkowej niezależności obydwu zmiennych. Niemniej test warunkowy porównujący

modele $[XY_{UA}][XZ_{UA}][YZ_{UA}]$ oraz $[XY_{UA}][YZ_{UA}]$, pokazuje, że należałoby odrzucić hipotezę $\delta_{ij}^{XZ} = 1$ ($G^2 = 76,11 - 28,47 = 47,64$ przy jednym stopniu swobody ($p < 0,0001$). Można oczekiwać, że zależność ta powinna być silniejsza, gdyż sytuacja materialna jest na ogół silnie skorelowana z wykształceniem. Należy jednak pamiętać, że pytamy respondentów o zadowolenie z sytuacji finansowej, a więc mamy do czynienia z subiektywną oceną, która nie musi silnie zależeć od wykształcenia.

Powstaje pytanie, czy dodanie kolejnych parametrów do modelu $[XY_{UA}][XZ_{UA}][YZ_{UA}]$ może poprawić dopasowanie do danych w sposób istotny. Można zapytać czy:

1. nie uzyskalibyśmy lepszego opisu gdybyśmy oszacowali w modelu logarytmiczno–multiplikatywnym odległości pomiędzy kategoriami zmiennych $X$, $Y$, $Z$,

2. uwzględnili bardziej skomplikowany typ związku aniżeli w modelu jednakowej interakcji, tj. model wierszowy (kolumnowy), bądź wierszowo–kolumnowy?

Aby uzyskać odpowiedź na pytanie pierwsze sformułowany został model: $[XY_{RC2}][XZ_{RC2}][YZ_{RC2}]$, przy czym przyjęto, że skalowanie danej zmiennej jest takie samo bez względu na to, której interakcji dotyczy np. skalowanie dla zmiennej $X$ jest takie samo w przypadku interakcji $[XY]$ oraz $[XZ]$. Model taki jest dobrze dopasowany do danych, a poszczególne parametry skalujące wynoszą odpowiednio: w przypadku zmiennej $X$: $u_1 = 0,53$, $u_2 = 0,41$, $u_3 = -0,23$, $u_4 = -0,70$; dla zmiennej $Y$: $v_1 = -0,72$, $v_2 = 0,02$, $v_3 = 0,69$; dla $Z$: $w_1 = -0,43$, $w_2 = -0,38$, $w_3 = 0,82$. Okazuje się, że największe odstępstwo w stosunku do założenia o jednakowych odległościach dotyczy dwóch pierwszych kategorii zmiennej $Z$: dla osób *zadowolonych* i *mniej więcej zadowolonych* - odległość jest relatywnie mniejsza. Wyniki estymacji wskazują również, że relatywnie „bliskie" są dwie pierwsze kategorie wykształcenia: podstawowe i niepełne średnie. Model ten jest dobrze dopasowany do danych. Test warunkowy na poziomie istotności nie nakazuje jednak odrzucania modelu prostszego ($G^2 = 28,5 - 18,7 = 9,8$ przy czterech stopniach swobody ($p = 0,0439$).

Aby odpowiedzieć na drugie z powyżej postawionych pytań przetestowany został model $[XY_{RC1}][XZ_{RC1}][YZ_{RC1}]$ zakładający efekt wierszowo–kolumnowy dla każdej pary zmiennych. Formalnie model ten można zdefiniować jako:

$$\begin{aligned} \Theta_{ij(k)}^{XY(Z)} &= \delta_{i\cdot}^{XY}\delta_{\cdot j}^{XY}, \\ \Theta_{i(j)k}^{X(Y)Z} &= \delta_{i\cdot}^{XZ}\delta_{\cdot k}^{XZ}, \\ \Theta_{(i)jk}^{(X)YZ} &= \delta_{j\cdot}^{YZ}\delta_{\cdot k}^{YZ}. \end{aligned} \tag{2.66}$$

Porównując statystyki dopasowania tego modelu i modelu prostszego $[XY_{UA}][XZ_{UA}][YZ_{UA}]$ można zauważyć, że dopasowanie tego modelu nie jest istotnie lepsze $G^2 = 14,47$ przy 8 stopniach swobody $p = 0,07$. Niemniej znaczna redukcja wielkości obydwu statystyk sugeruje, że uwzględnienie bardziej skomplikowanego związku niż model jednakowej interakcji może być uzasadnione, choć należałoby wybrać model prostszy niż rozpatrywany[16].

Aby to sprawdzić, formułujemy kolejne hipotezy. W każdym z modeli $[XY_{RC1}][XZ_{UA}][YZ_{UA}]$, $[XY_{UA}][XZ_{RC1}][YZ_{UA}]$, $[XY_{UA}][XZ_{UA}][YZ_{RC1}]$ wprowadzany jest efekt wierszowo–kolumnowy w odniesieniu do jednej pary zmiennych. Porównanie modelu $[XY_{UA}][XZ_{UA}][YZ_{RC1}]$ oraz zagnieżdżonego w nim modelu $[XY_{UA}][XY_{UA}][YZ_{UA}]$ pokazuje, że nie różnią się one w sposób istotny statystycznie. Kierując się tym kryterium pozostaniemy przy hipotezie o jednakowej interakcji w odniesieniu do zmiennych $Y$ oraz $Z$. Innymi słowy, zakładamy, że $\delta_{j\cdot}^{YZ} = const$ oraz $\delta_{\cdot k}^{YZ} = const$.

Odnośnie interakcji dla pary zmiennych $X$ oraz $Z$ porównanie modelu $[XY_{UA}][XZ_{RC1}][YZ_{UA}]$ z modelem prostszym – analogicznie jak wyżej – wskazuje, że uwzględnienie bardziej złożonego wzoru zależności może być uzasadnione. Założenie $\delta_{i\cdot}^{XZ} = const$ oraz $\delta_{\cdot k}^{XZ} = const$ można odrzucić na przyjętym przez nas poziomie istotności równym $\alpha = 0,05$. Test warunkowy posiada trzy stopnie swobody, $G^2 = 8,7$ $p = 0,03$. Wynik ten sugeruje, że jednakowa interakcja może być założeniem zbyt prostym. Zanim przyjmiemy efekt wierszowo–kolumnowy warto zapytać, czy nie jest możliwe przyjęcie hipotezy prostszej: o efekcie wierszowym lub efekcie kolumnowym. Okazuje się, że o ile przyjęcie modelu wierszowego (tj. związanego ze zmienną $X$) nie prowadzi do znaczącej redukcji statystyki $G^2$, to można wziąć pod uwagę przyjęcie modelu kolumnowego, w którym uwzględniamy specyfikę zmiennej $Z$, opisującej ocenę własnej sytuacji materialnej. Porównanie modeli $[XZ_{UA}][XZ_C][YZ_{UA}]$ oraz $[XY_{UA}][XZ_{UA}][YZ_{UA}]$ stanowi warunkowy test dla założenia $\delta_{i\cdot}^{XZ} = const$: $G^2 = 5,23$, $df = 1$ $p = 0,02$.

Postępując analogicznie – porównanie modeli $[XY_{RC1}][XZ_{UA}][YZ_{UA}]$ oraz $[XY_{UA}][XZ_{UA}][YZ_{UA}]$ stanowi warunkowy test dla założenia $\delta_{i\cdot}^{XY} = const$ oraz $\delta_{\cdot j}^{XY} = const$. Hipotezy tej nie można odrzucić na poziomie istotności równym 0,05, niemniej redukcja statystyki $G^2$ jest na tyle znacząca, że warto przetestować, efekt kolumnowy i efekt wierszowy w odniesieniu do interakcji pomiędzy $X$ oraz $Y$. Testy warunkowe dla efektu wierszowego (czyli wykształcenia) jak i efektu kolumnowego

[16]Porównanie modeli $[XY_{RC1}][XZ_{RC1}][YZ_{RC1}]$ oraz $[XY][XZ][YZ]$ pokazuje, że redukcja statystyki $G^2$ jest niewielka, co sugeruje, że dla żadnej pary zmiennych nie jest celowe uwzględnienie interakcji o nieokreślonym wzorze zależności (FA), dlatego w dalszej części analizy danych z tabeli 2.20 nie będą formułowane tego typu modele.

(związanego ze zmienną $Y$) względem hipotezy o jednakowej interakcji nie prowadzą do odrzucenia hipotezy prostszej. Porównanie modeli $[XY_R][XZ_{UA}][YZ_{UA}]$ oraz $[XY_{UA}][XZ_{UA}][YZ_{UA}]$ daje wartość $G^2 = 4,66$, $df = 2$ $p = 0,09$, a dla analogicznego testu dla modelu $[XY_C][XZ_{UA}][YZ_{UA}]$ otrzymujemy $G^2 = 1,1$, $df = 1$ $p = 0,29$. Te testy pokazują, że uwzględnienie efektu kolumnowego wydaje się zbędne, podobnie nie wydaje się konieczne podobnie przyjęcie efektu wierszowego, warto jednak odnotować, że w tym przypadku redukcja statystyki $G^2$ jest większa.

Powyższe analizy sugerują wybranie modelu $[XY_{UA}][XZ_C][YZ_{UA}]$ uwzględniającego w interakcji pomiędzy $X$ oraz $Z$ specyfikę zmiennej opisującej opinie dotyczące dochodów. Formalnie:

$$\begin{aligned} \Theta_{ij(k)}^{XY(Z)} &= \delta^{XY}, \\ \Theta_{i(j)k}^{X(Y)Z} &= \delta_{\cdot k}^{XZ}, \\ \Theta_{(i)jk}^{(X)YZ} &= \delta^{YZ}. \end{aligned} \tag{2.67}$$

Zgodnie z tym modelem:

1. Związek pomiędzy wykształceniem ($X$) i opiniami dotyczącymi zmniejszenia zróżnicowania dochodów przez rząd ($Y$) daje się opisać za pomocą jednego parametru $\delta^{XY} = 1,47$. Parametr ten opisuje wszystkie warunkowe lokalne stosunki szans, które są sobie równe:

   $$\Theta_{11(k)}^{XY(Z)} = \Theta_{12(k)}^{XY(Z)} = \Theta_{21(k)}^{XY(Z)} = \Theta_{22(k)}^{XY(Z)} = 1,47.$$

   Co więcej, ich wartość nie zależy od tego, dla której podzbiorowości $Z = z_k$ są one wyznaczone. Dla przykładu: proporcja osób, które zgadzają się co do tego, że rząd powinien niwelować zróżnicowanie w dochodach do osób, które w tej kwestii są neutralne jest prawie 1,5 razy większa wsród osób z wykształceniem wyższym aniżeli w grupie osób z wykształceniem średnim. Podobnie można opisać związek pomiędzy opiniami respondentów ($Y$) i oceną własnej sytuacji materialnej ($Z$) przy czym lokalny stosunek szans wynosi 0,69.

2. Związek między oceną sytuacji materialnej ($Z$) i opiniami dotyczącymi działań rządu ($Y$) jest opisywany za pomocą efektu kolumnowego, związanego ze zmienną $Z$. Odpowiednie parametry kolumnowe są równe $\phi_1 = 1$, $\phi_2 = 0,98$, $\phi_3 = 0,77$. Te parametry pokazują, że lokalne stosunki szans wyznaczone dla dwóch pierwszych kategorii zmiennej $Z$ są bliskie jedności. Oznacza to, że rozkłady warunkowe $Y$ są zbliżone wśród osób *zadowolonych* i *mniej więcej zadowolonych* z sytuacji materialnej. Inaczej wygląda to, jeśli porówna się dwie

wyżej wymienione kategorie z osobami *niezadowolonymi*. W tej ostatniej relatywnie częściej można spotykać osoby o postawach egalitarnych, tj. postulujące zmniejszenie różnic w dochodach.

Oczywiście, możliwe jest sformułowanie wielu innych modeli. Na przykład, w odniesieniu do interakcji pomiędzy $X$ i $Y$ zamieszczone powyżej testy warunkowe sugerowały, że uwzględnienie zmiennej wierszowej ($X$), przynosi znaczącą — choć nieistotną statystycznie — redukcję statystyki $G^2$. Moglibyśmy rozważyć model $[XY_R][XZ_C][YZ_{UA}]$. Test warunkowy porównujący ten model z rozważanym powyżej modelem $[XY_{UA}][XZ_C][YZ_{UA}]$ nie prowadzi do odrzucenia modelu prostszego: $G^2 = 4,82$, $df = 2$ ($p = 0,09$). Wydaje się, więc ze dalsze modyfikowanie modelu $[XY_{UA}][XZ_C][YZ_{UA}]$ przez wprowadzanie do niego kolejnych parametrów nie jest celowe.

Trzeba podkreślić, że model ten wybraliśmy przyjmując przy weryfikacji hipotez test istotności równy 0,05. Wybierając niższy poziom istotności, np. 0,01 wybralibyśmy prostszy model $[XY_{UA}][XZ_{UA}][YZ_{UA}]$[17]. Powyższy przykład służył prezentacji różnych modeli dla zmiennych porządkowych. Pamiętać jednak należy, że przy weryfikacji hipotez istotne są również przesłanki teoretyczne i wiedza badacza. Jeśli prowadzą one do sformułowania modelu, o dogodnej interpretacji, a jednocześnie weryfikacja tego modelu nie zmusza nas do jego odrzucenia, nie zawsze wydaje się konieczne wprowadzanie do modelu kolejnych parametrów, nawet jeśli poprawiają one „dopasowanie" do danych. Przesłanki teoretyczne, mogą również wskazywać, które modele powinno się ze sobą porównywać.

### 2.4.2 Modelowanie interakcji trzeciego rzędu

W poprzedniej części omówione zostały modele dla trzech zmiennych, w których związek pomiędzy dwiema zmiennymi nie zależał od wartości trzeciej zmiennej, innymi słowy spełniony był warunek 2.57. W tej części zaprezentujemy modele, które nie czynią takiego założenia, np. lokalny stosunek szans opisujący warunkowy związek pomiędzy $X$ i $Y$, tj:

$$\Theta_{ij(k)}^{XY(Z)} = \frac{\pi_{ijk}^{XYZ}\pi_{(i+1)(j+1)k}^{XYZ}}{\pi_{(i+1)jk}^{XYZ}\pi_{i(j+1)k}^{XYZ}}, \tag{2.68}$$

zależy od tego, dla której wartkości trzeciej zmiennej został wyróżniony (inaczej było w formule (2.57), która opisuje hipotezę $[XY][XZ][YZ]$). Dla przykładu związek pomiędzy zadowoleniem z życia i oceną własnej sytuacji materialnej może zależeć od

[17] Również posługiwanie się indeksem BIC, który został omówiony w pierwszym rozdziale prowadzi do wybrania tego modelu.

wykształcenia badanych osób. Podobnie można modelować wielkości $\Theta_{i(j)k}^{X(Y)Z}$, $\Theta_{(i)jk}^{(X)YZ}$. Warto już w tym miejscu podkreślić, że wielkości te są od siebie w dużym zakresie zależne, tj. przyjęcie hipotezy odnośnie warunkowego stosunku szans $\Theta_{ij(k)}^{XY(Z)}$ pośrednio nakłada też warunki dotyczące wielkości $\Theta_{i(j)k}^{X(Y)Z}$ i $\Theta_{(i)jk}^{(X)YZ}$[18].

Interakcję trzeciego rzędu można również modelować odwołując się do relacji pomiędzy lokalnymi stosunkami szans dla zmiennych $X$ i $Y$ wyróżnionymi dla sąsiednich kategorii zmiennej $Z$, tj:

$$\begin{aligned}\Theta_{ijk}^{XYZ} &= \frac{\Theta_{ij(k+1)}^{XY(Z)}}{\Theta_{ij(k)}^{XY(Z)}} = \frac{\Theta_{i(j+1)k}^{X(Y)Z}}{\Theta_{i(j)k}^{X(Y)Z}} = \frac{\Theta_{(i+1)jk}^{(X)YZ}}{\Theta_{(i)jk}^{(X)YZ}} = \\ &= \frac{(\pi_{ijk}^{XYZ}\pi_{(i+1)(j+1)k}^{XYZ})/(\pi_{(i+1)jk}^{XYZ}\pi_{i(j+1)k}^{XYZ})}{(\pi_{ij(k+1)}^{XYZ}\pi_{(i+1)(j+1)(k+1)}^{XYZ})/(\pi_{(i+1)j(k+1)}^{XYZ}\pi_{i(j+1)(k+1)}^{XYZ})}.\end{aligned} \tag{2.69}$$

Wielkość ta — nazywana *lokalnym stosunkiem szans trzech zmiennych* — opisuje związek pomiędzy wybranymi kategoriami trzech zmiennych. Z powyższego zapisu wynika, że $\Theta_{ijk}^{XYZ}$ jest wielkością symetryczną, tj. opisuje również ilorazy stosunków szans dla $X$ i $Z$ wyróżnionych względem $Y$, jak również $Y$ i $Z$ wyróżnionymi względem $X$. W odniesieniu do przykładu podanego powyżej wielkość ta może określać na ile różni się związek pomiędzy zadowoleniem z życia i oceną własnej sytuacji materialnej, jeśli porównuje się kolejne kategorie wykształcenia.

Na ogół tę samą hipotezę można sformułować zarówno posługując się warunkowymi stosunkami szans (2.68), jak też lokalnym stosunkiem szans trzech zmiennych (2.69). Używanie jednego bądź drugiego sposobu może być jednak mniej lub bardziej wygodne w zależności od przedmiotu zainteresowania badacza. Na przykład, jeśli w centrum zainteresowania badacza jest związek pomiędzy wybraną parą zmiennych a trzecia jest raczej zmienną kontrolną bądź grupującą bardziej adekwatne wydaje się odwołanie do warunkowych stosunków szans. Odnosząc się do podanego powyżej przykładu zmienną wykształcenie badacz może traktować jako zmienną kontrolną: interesuje go związek pomiędzy zadowoleniem z życia ($X$) i sytuacją materialną ($Y$), ale jednocześnie przypuszcza, że siła tego związku zależy od poziomu wykształcenia ($Z$), dlatego należy uwzględnić tę zmienną w modelu. Przyjęta hipoteza odnośnie zmiennych $X$ i $Y$ względem $Z$ może *implicite* zakładać również jakiś rodzaj związku pomiędzy wykształceniem a zadowoleniem z życia, mimo, że nie jest on w centrum zainteresowania badacza. Podobnie, jeśli mamy trzy zmienne: kraj, wykształcenie,

[18]Inaczej było w sytuacji, gdy zakładaliśmy, że nie występuje interakcja trzeciego rzędu: jak pamiętamy na ogół możliwe było niezależne modelowanie interakcji drugiego rzędu dla poszczególnych par zmiennych. Na przykład, jeśli założymy, że warunkowy związek pomiędzy $X$ oraz $Y$ daje się opisać za pomocą jednakowej interakcji, tj. $\Theta_{ij(k)}^{XY(Z)} = \delta$, nie wykluczało to, że zmienne $Y$ oraz $Z$ są warunkowo niezależne tj. $\Theta_{(i)jk}^{(X)YZ} = 1$.

opinie dotyczące zmniejszenia zróżnicowania dochodów, badacz może być zainteresowany odpowiedzią na pytanie czy związek pomiędzy dwiema ostatnimi zmiennymi jest taki sam, czy też różny w poszczególnych krajach. Możliwe jest jednak, że badacza interesuje związek pomiędzy trzema zmiennymi (np. zadowolenie z życia, sytuacja materialna i wykształcenie) i z perspektywy założeń badawczych żadna z par zmiennych nie wydaje się ważniejsza. Wówczas bardziej adekwatne może wydać się formułowanie hipotez za pomocą lokalnego stosunku szans trzech zmiennych. Zaprezentujemy obydwa podejścia, tj.

- modele współzależności trzech zmiennych definiowane na ogół za pomocą lokalnego stosunku szans trzech zmiennych,
- modele warunkowej zależności dwóch zmiennych formułowane za pomocą warunkowych stosunków szans.

Prezentując konkretne modele niejednokrotnie będą prezentowane obydwie formuły, pokazując tym samym, że w wielu sytuacjach powyższe rozróżnienie ma głównie znaczenie interpretacyjne.

Zanim omawiane będą poszczególne modele dotyczące interakcji trzeciego rzędu, warto zauważyć, że nie muszą być one bardziej złożone od hipotezy, zakładającej równość warunkowych stosunków szans $[XY][XZ][YZ]$. Co prawda, modelując interakcję trzeciego rzędu nie zakładamy, że $d_{ijk}^{XYZ} = 1$, to jednocześnie korzystając z faktu, że kategorie zmiennych są uporządkowane możemy modelować interakcję drugiego rzędu w sposób prostszy niż w modelu $[XY][XZ][YZ]$. Oznacza, to, że część modeli jest nieporównywalna z modelem równości warunkowych stosunków szans pod względem prostoty. Jak zobaczymy w dalszej części, wiele z modeli, które przedstawimy będzie modelowało łącznie interakcję drugiego i trzeciego rzędu, to znaczy w odniesieniu do parametrów modelu logarytmiczno–liniowego hipoteza dotyczyć będzie iloczynu $d_{ij}^{XY} \cdot d_{ijk}^{XYZ}$.

### Przykłady modeli współzależności trzech zmiennych

W tej części podanych zostanie kilka przykładów modeli formułowanych dla trzech zmiennych porządkowych. Pierwsza z prezentowanych hipotez $[XYZ_U]$ zakładać będzie, że lokalny stosunek szans trzech zmiennych jest taki sam dla każdej kombinacji wartości trzech zmiennych, tj:

$$\Theta_{ijk}^{XYZ} = \delta \tag{2.70}$$

Ilustracją tej hipotezy jest tabela 2.23, która uwzględnia parametry drugiego i trzeciego rzędu. Jak widać dowolny iloraz warunkowych lokalnych stosunków szans

Tabela 2.23: Ilustracja hipotezy $[XYZ_U]$ — parametry interakcji

| $Z = z_1$ | | | | |
|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
| $x_1$ | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\gamma_{22}^{XY}$ | $\gamma_{23}^{XY}$ | $\gamma_{24}^{XY}$ |
| $x_3$ | 1 | $\gamma_{32}^{XY}$ | $\gamma_{33}^{XY}$ | $\gamma_{34}^{XY}$ |
| $x_4$ | 1 | $\gamma_{42}^{XY}$ | $\gamma_{43}^{XY}$ | $\gamma_{44}^{XY}$ |
| $Z = z_2$ | | | | |
| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
| $x_1$ | 1 | $\gamma_{22}^{YZ}$ | $\gamma_{32}^{YZ}$ | $\gamma_{42}^{YZ}$ |
| $x_2$ | $\gamma_{22}^{XZ}$ | $\gamma_{22}^{XY}\gamma_{22}^{XZ}\gamma_{22}^{YZ}\delta$ | $\gamma_{23}^{XY}\gamma_{22}^{XZ}\gamma_{32}^{YZ}\delta^2$ | $\gamma_{24}^{XY}\gamma_{22}^{XZ}\gamma_{42}^{YZ}\delta^3$ |
| $x_3$ | $\gamma_{32}^{XZ}$ | $\gamma_{32}^{XY}\gamma_{32}^{XZ}\gamma_{22}^{YZ}\delta^2$ | $\gamma_{33}^{XY}\gamma_{32}^{XZ}\gamma_{32}^{YZ}\delta^4$ | $\gamma_{34}^{XY}\gamma_{32}^{XZ}\gamma_{42}^{YZ}\delta^6$ |
| $x_4$ | $\gamma_{42}^{XZ}$ | $\gamma_{42}^{XY}\gamma_{42}^{XZ}\gamma_{22}^{YZ}\delta^3$ | $\gamma_{43}^{XY}\gamma_{42}^{XZ}\gamma_{32}^{YZ}\delta^6$ | $\gamma_{44}^{XY}\gamma_{42}^{XZ}\gamma_{42}^{YZ}\delta^9$ |
| $Z = z_3$ | | | | |
| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
| $x_1$ | 1 | $\gamma_{23}^{YZ}$ | $\gamma_{33}^{YZ}$ | $\gamma_{43}^{YZ}$ |
| $x_2$ | $\gamma_{23}^{XZ}$ | $\gamma_{22}^{XY}\gamma_{23}^{XZ}\gamma_{23}^{YZ}\delta^2$ | $\gamma_{23}^{XY}\gamma_{23}^{XZ}\gamma_{33}^{YZ}\delta^4$ | $\gamma_{24}^{XY}\gamma_{23}^{XZ}\gamma_{43}^{YZ}\delta^6$ |
| $x_3$ | $\gamma_{33}^{XZ}$ | $\gamma_{32}^{XY}\gamma_{33}^{XZ}\gamma_{23}^{YZ}\delta^4$ | $\gamma_{33}^{XY}\gamma_{33}^{XZ}\gamma_{33}^{YZ}\delta^8$ | $\gamma_{34}^{XY}\gamma_{33}^{XZ}\gamma_{43}^{YZ}\delta^{12}$ |
| $x_4$ | $\gamma_{43}^{XZ}$ | $\gamma_{42}^{XY}\gamma_{43}^{XZ}\gamma_{23}^{YZ}\delta^6$ | $\gamma_{43}^{XY}\gamma_{43}^{XZ}\gamma_{33}^{YZ}\delta^{12}$ | $\gamma_{44}^{XY}\gamma_{43}^{XZ}\gamma_{43}^{YZ}\delta^{18}$ |

dotyczący dwóch zmiennych np. $X$ i $Y$ wyróżniony w dwóch sąsiednich kategoriach trzeciej zmiennej $Z$ jest równy $\delta$. Na przykład:

$$\begin{aligned}\Theta_{132}^{XYZ} &= \frac{\Theta_{(2)32}^{(X)YZ}}{\Theta_{(1)32}^{(X)YZ}} = \frac{\Theta_{1(4)2}^{X(Y)Z}}{\Theta_{1(3)2}^{X(Y)Z}} = \frac{\Theta_{13(3)}^{XY(Z)}}{\Theta_{13(2)}^{XY(Z)}} = \\ &= \frac{(\gamma_{33}^{YZ}\gamma_{24}^{XY}\gamma_{23}^{XZ}\gamma_{43}^{YZ}\delta^6)/(\gamma_{43}^{YZ}\gamma_{23}^{XY}\gamma_{23}^{XZ}\gamma_{33}^{YZ}\delta^4)}{(\gamma_{32}^{YZ}\gamma_{24}^{XY}\gamma_{22}^{XZ}\gamma_{42}^{YZ}\delta^3)/(\gamma_{42}^{YZ}\gamma_{23}^{XY}\gamma_{22}^{XZ}\gamma_{32}^{YZ}\delta^2)} = \delta.\end{aligned}$$

Ilustracja w tabeli 2.23 jest zgodna z parametryzacją względem kategorii odniesienia $x_1$, $y_1$, $z_1$. Bardziej ogólnie model ten można przedstawić jako:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ij}^{XY} \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \delta^{(i-a)(j-b)(k-c)}, \tag{2.71}$$

gdzie kategoriami odniesienia są odpowiednio $x_a$, $y_b$, $z_c$. Model nasycony posiada (r-1)(c-1)(t-1) parametrów interakcji trzeciego rzędu, natomiast omawiany model równości lokalnego stosunku szans trzech zmiennych, posiada tylko jeden parametr tego typu. Liczba stopni swobody wynosi więc $(r-1)(c-1)(t-1)-1$.

W tabeli 2.24 przedstawiony został rozkład trzech zmiennych opisujących opinie dotyczące trzech kwestii: tego, czy rząd powinien angażować się w gospodarkę ($X$), czy rząd powinien zmniejszyć różnice w dochodach ($Y$) oraz tego, czy pracownicy dla ochrony warunków pracy i wysokości płac potrzebują silnych związków zawodowych ($Z$). Dane pochodzą z polskiej edycji I tury Europejskiego Sondażu Społecznego (2002). [19].

Tabela 2.24: Opinie dotyczące działań rządu i roli związków zawodowych[a]

| Przeciwnicy angażowania się rządu w gospodarkę ($Z = 1$) | | | |
|---|---|---|---|
| Rząd powinien zmniejszyć różnice w dochodach ($Y$) | Pracownicy potrzebują silnych związków zawodowych ($X$) | | |
| | 1. Zgadzam się | 2. Ani się zgadzam, ani nie zgadzam | 3. Nie zgadzam się |
| 1. Zgadzam się | 310,5 | 41,5 | 15,0 |
| 2. Ani się zgadzam ani nie zgadzam | 24,0 | 10,4 | 8,1 |
| 3. Nie zgadzam się | 34,0 | 19,1 | 24,1 |
| Osoby neutralne w kwestii angażowania się rządu w gospodarkę ($Z = 2$) | | | |
| 1. Zgadzam się | 315,1 | 32,5 | 13,0 |
| 2. Ani się zgadzam ani nie zgadzam | 33,2 | 36,0 | 4,8 |
| 3. Nie zgadzam się | 20,4 | 2,9 | 3,4 |
| Zwolennicy angażowania się rządu w gospodarkę ($Z = 3$) | | | |
| 1. Zgadzam się | 670,0 | 47,8 | 28,9 |
| 2. Ani się zgadzam ani nie zgadzam | 49,3 | 8,2 | 4,8 |
| 3. Nie zgadzam się | 71,6 | 12,6 | 16,0 |

[a]Źródło: Europejski Sondaż Społeczny, 2002. Dane przeważone.

[19]Respondenci mieli ustosunkować się do stwierdzeń: 1.*Im mniej rząd angażuje się w sferę gospodarczą, tym lepiej dla Polski*, 2. *Rząd powinien podjąć działania zmierzające do zmniejszenia różnic w dochodach*, 3. *Pracownicy potrzebują silnych związków zawodowych dla ochrony warunków pracy i poziomu płac*, na pięciopunktowej skali *1. Zdecydowanie się zgadzam, 2. Zgadzam się, 3. Ani się zgadzam, ani nie zgadzam, 4. Nie zgadzam się, 5. Zdecydowanie się nie zgadzam*. Ze względu na konieczność zapewnienia odpowiednich liczebności komórkom rozkładu łącznego, połączone zostały kategorie 1 i 2 oraz kategorie 4 i 5. Osoby, które nie udzieliły odpowiedzi na przynajmniej jedno z pytań zostały wyłączone z analizy (stanowiły one 11,5% wszystkich respondentów).

W odniesieniu do tych danych można odrzucić hipotezę o równości warunkowych stosunków szans $\chi^2 = 21,4$ ($p = 0,006$), $G^2 = 22,06$ ($p = 0,004$) $df = 8$. Oznacza, to, że związek pomiędzy dowolną parą zmiennych różni się ze względu na wartości trzeciej zmiennej, tj. występuje interakcja trzeciego rzędu. Okazuje się, że interakcja ta może być opisana za pomocą jednego parametru. Hipoteza głosząca równość lokalnych stosunków szans trzech zmiennych jest akceptowalna na poziomie istotności równym $\chi^2 = 16,4$ ($p = 0,06$), $G^2 = 16,90$ ($p = 0,05$) $df = 9$. Parametr opisujący lokalny stosunek szans trzech zmiennych wynosi $\Theta_{ijk}^{XYZ} = \delta = 0,88$. Wielkość ta zostanie zinterpretowana przykładowo dla stosunku szans wyróżnionego dla pierwszych dwóch kategorii zmiennych $X$ oraz $Y$. W podzbiorowości $Z = 1$, warunkowy stosunek szans wynosi $\Theta_{11(1)}^{XY(Z)} = 1,96$. Oznacza to, że zgodnie z powyższym modelem wśród osób uważających, że *im mniej rząd angażuje się w gospodarkę tym lepiej dla Polski*, proporcja osób zgadzających się co do tego, że pracownicy potrzebują silnych związków zawodowych do osób neutralnych w tej kwestii, jest prawie dwukrotnie większa wśród osób uważających, że rząd powinien podjąć działania zmierzające do zmniejszenia różnic w dochodach niż wśród osób, które z tym stwierdzeniem*ani się zgadzają, ani się nie zgadzają*. Wielkość $\delta$ wskazuje, że analogiczny stosunek szans jest 1,14 (tj. 1/0,88) razy mniejszy wśród osób, które pozostają neutralne w kwestii oceny skutków angażowania się rządu w gospodarkę, (tj. $\Theta_{11(2)}^{XY(Z)} = 1,96 \cdot 0,88 = 1,72$), a wśród osób, które nie zgadają się co do tego, że *im mniej rząd angażuje się w gospodarkę tym lepiej* jest 1,30 (czyli $1/0,88^2$) razy mniejszy i wynosi $\Theta_{11(3)}^{XY(Z)} = 1,50$.

Rysunek 2.5 pokazuje graficzną interpretację tego modelu. Jak widać logarytmy warunkowych stosunków szans wyróżnionych dla kolejnych wartości zmiennej $Z$ leżą na jednej linii a parametr nachylenia tej linii, tj. różnica dla dwóch kolejnych wartości zmiennej $Z$ jest równa logarytmowi lokalnego stosunku szans trzech zmiennych:

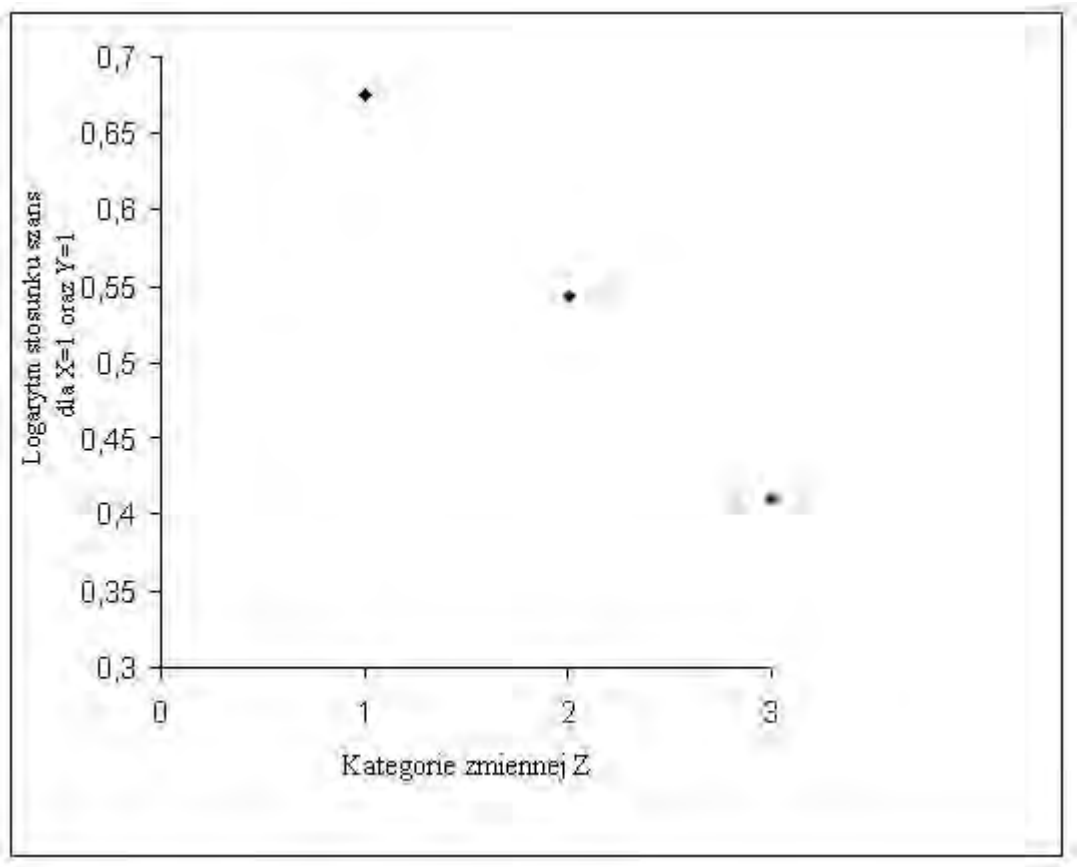
$$\log \Theta_{11(k+1)}^{XY(Z)} - \log \Theta_{11(k)}^{XY(Z)} = \log \frac{\Theta_{11(k+1)}^{XY(Z)}}{\Theta_{11(k)}^{XY(Z)}} = \log \delta.$$

Powyższa interpretacja dotyczy stosunku szans dla dwóch pierwszych kategorii zmiennych $X$ oraz $Y$, jednak w podobnej relacji pozostają warunkowe stosunki szans jakie można wyróżnić dla dowolnych innych kategorii zmiennych $X$ oraz $Y$, np. $\Theta_{21(2)}^{XY(Z)} / \Theta_{21(1)}^{XY(Z)} = \delta = 0,88$, itd.

Jak pokazywała formuła 2.69, lokalny stosunek szans trzech zmiennych jest wielkością symetryczną, tak więc w podobnej relacji pozostają stosunki szans jakie możemy wyróżnić dla zmiennych $X$ i $Z$ względem $Y$ oraz $Y$ i $Z$ względem $X$, tj.

$$\frac{\Theta_{i(j+1)k}^{X(Y)Z}}{\Theta_{i(j)k}^{X(Y)Z}} = \frac{\Theta_{(i+1)jk}^{(X)YZ}}{\Theta_{(i)jk}^{(X)YZ}} = \delta = 0,88,$$

Rysunek 2.5: Warunkowe logarytmy lokalnych stosunków szans dla pierwszych kategorii zmiennych $X$ oraz $Y$ wyróżnione dla kolejnych kategorii zmiennej $Z$



dla dowolnych $i = 1, \ldots, r-1$, $j = 1, \ldots, c-1$, $k = 1, \ldots, t-1$.

Hipotezę tę można wyrazić również odwołując się do warunkowych stosunków szans. Zgodnie z tym ujęciem spełnione są następujące warunki:

$$\begin{aligned} \Theta_{ij(k)}^{XY(Z)} &= \delta_{ij}^{XY}\delta^{k}, \\ \Theta_{i(j)k}^{X(Y)Z} &= \delta_{ik}^{XZ}\delta^{j}, \\ \Theta_{(i)jk}^{(X)YZ} &= \delta_{jk}^{YZ}\delta^{i}. \end{aligned} \tag{2.72}$$

Jak widać, każdy stosunek szans jest definiowany przez parametr specyficzny dla każdej pary zmiennych, co więcej dla każdej kombinacji wartości zmiennych: parametry te — $\delta_{ij}^{XY}$, $\delta_{ik}^{XZ}$, $\delta_{jk}^{YZ}$ — opisują interakcję drugiego rzędu. Natomiast parametr $\delta$, definiujący interakcję trzeciego rzędu, jest identyczny dla każdej pary zmiennych.

Powyższa hipoteza nie zakłada niczego odnośnie interakcji drugiego rzędu. Dla każdej pary zmiennych wzór zależności jest nieokreślony (FA). Model ten można uprościć formułując dodatkowe warunki definiujące kształt tej zależności, dla jednej pary zmiennych bądź wszystkich par. Przypuśćmy, że chcemy założyć dodatkowo, że interakcja drugiego rzędu dla każdej pary zmiennych może być opisana za pomocą modelu jednakowej interakcji. Formalnie:

$$\begin{aligned} \Theta_{ij(k)}^{XY(Z)} &= \delta^{XY}\delta^{k}, \\ \Theta_{i(j)k}^{X(Y)Z} &= \delta^{XZ}\delta^{j}, \\ \Theta_{(i)jk}^{(X)YZ} &= \delta^{YZ}\delta^{i}. \end{aligned} \tag{2.73}$$

Tej hipotezy, którą oznaczać będziemy jako $[XY_{UA}][XZ_{UA}][YZ_{UA}][XYZ_{U}]$ nie możemy sformułować wyłącznie za pomocą lokalnego stosunku szans trzech zmiennych, gdyż nie dotyczy ona jedynie interakcji trzeciego rzędu. Zauważmy, że w odróżnieniu od poprzedniej hipotezy (2.72) parametry interakcji drugiego rzędu w formule 2.73 nie są indeksowane przez kategorie poszczególnych zmiennych. Parametryzacja tej hipotezy przedstawia się następująco:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \left(\delta^{XY}\right)^{(i-a)(j-b)} \cdot \left(\delta^{XZ}\right)^{(i-a)(k-c)} \cdot \left(\delta^{YZ}\right)^{(j-b)(k-c)} \delta^{(i-a)(j-b)(k-c)}, \tag{2.74}$$

gdzie kategoriami odniesienia są $x_a$, $y_b$, $z_c$. Model ten — pomimo, że opisuje interakcję drugiego i trzeciego rzędu — wykorzystuje jedynie cztery parametry więcej niż model niezależności stochastycznej trzech zmiennych. Liczba stopni swobody tego modelu wynosi $df = rcl - r - c - t - 2$.

W obydwu prezentowanych powyżej hipotezach, interakcję trzeciego rzędu można opisać za pomocą jednego parametru. Jak pokazuje rysunek 2.5 siła związku pomiędzy dwiema zmiennymi rosła bądź malała proporcjonalnie dla kolejnych kategorii trzeciej

zmiennej. Możliwe jest jednak modelowanie interakcji trzeciego rzędu w sposób bardziej złożony np. po uwzględnieniu odległości pomiędzy kolejnymi wartościami zmiennych porządkowych, podobnie jak robiliśmy to dla logarytmiczno–multiplikatywnego modelu wierszowo–kolumnowego. Hipotezę taką można sformułować następująco:

$$\Theta_{ijk}^{XYZ} = \delta^{(u_{i+1}-u_i)(v_{j+1}-v_j)(w_{k+1}-w_k)}, \tag{2.75}$$

gdzie parametry $u_i$, $v_j$, $w_k$, są parametrami szacowanymi w modelu. Warunkowe stosunki szans można sformułować dla tej hipotezy w sposób następujący:

$$\begin{aligned} \Theta_{ij(k)}^{XY(Z)} &= \delta_{ij}^{XY}\delta^{(u_{i+1}-u_i)(v_{j+1}-v_j)w_k}, \\ \Theta_{i(j)k}^{X(Y)Z} &= \delta_{ik}^{XZ}\delta^{(u_{i+1}-u_i)(w_{k+1}-w_k)v_j}, \\ \Theta_{(i)jk}^{(X)YZ} &= \delta_{jk}^{YZ}\delta^{(v_{j+1}-v_j)(w_{k+1}-w_k)u_i}. \end{aligned} \tag{2.76}$$

Hipotezę tę można parametryzować w następujący sposób:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ij}^{XY} \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \delta^{(u_{i+1}-u_i)(v_{j+1}-v_j)(w_{k+1}-w_k)}. \tag{2.77}$$

Na wielkości tych parametrów $u_i$, $v_j$, $w_k$ można nałożyć warunki analogiczne do 2.47. Jeśli chodzi o liczbę niezależnych parametrów opisujących interakcję trzeciego rzędu, model posiada parametr $\delta$, $(r-2)$ parametrów $u_i$, $(c-2)$ parametrów $v_j$, $(t-2)$ parametrów $w_k$. Jego liczba stopni swobody wynosi:

$$df = (r-1)(c-1)(t-1) - (r-2) - (c-2) - (t-2) - 1.$$

Powyższy model logarytmiczno–multiplikatywny może być stosowany również do analizy związków pomiędzy trzema zmiennymi nominalnymi.

### Modelowanie warunkowej zależności dwóch zmiennych

Zaprezentowanych zostanie teraz kilka hipotez koncentrujących się na warunkowej zależności dwóch zmiennych w podzbiorowościach wyróżnionych ze względu na wartości trzeciej zmiennej. Na początek przedstawiona zostanie hipoteza *heterogenicznej jednakowej interakcji*, która głosi, że związek pomiędzy zmiennymi $X$ oraz $Y$ daje się opisać za pomocą hipotezy jednakowej interakcji, przy czym siła tej interakcji jest inna w każdej podzbiorowości wyróżnionej ze względu na zmienną $Z$. Hipoteza ta — którą oznaczać będziemy $[XY_{UA|Z}][XZ][YZ]$— głosi, że:

$$\Theta_{ij(k)}^{XY(Z)} = \delta_k. \tag{2.78}$$

Tabela 2.25 stanowi ilustrację tej hipotezy. Parametryzacja dla niej przedstawia się następująco:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \delta_k^{(i-a)(j-b)}, \tag{2.79}$$

Tabela 2.25: Ilustracja hipotezy $[XY_{UA|Z}][XZ][YZ]$ — parametry interakcji

| $Z = z_1$ | | | | |
|---|---|---|---|---|
| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
| $x_1$ | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\delta_1$ | $\delta_1^2$ | $\delta_1^3$ |
| $x_3$ | 1 | $\delta_1^2$ | $\delta_1^4$ | $\delta_1^6$ |
| $x_4$ | 1 | $\delta_1^3$ | $\delta_1^6$ | $\delta_1^9$ |
| $Z = z_2$ | | | | |
| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
| $x_1$ | 1 | $\gamma_{22}^{YZ}$ | $\gamma_{32}^{YZ}$ | $\gamma_{42}^{YZ}$ |
| $x_2$ | $\gamma_{22}^{XZ}$ | $\delta_2\gamma_{22}^{XZ}\gamma_{22}^{YZ}$ | $\delta_2^2\gamma_{22}^{XZ}\gamma_{32}^{YZ}$ | $\delta_2^3\gamma_{22}^{XZ}\gamma_{42}^{YZ}$ |
| $x_3$ | $\gamma_{32}^{XZ}$ | $\delta_2^2\gamma_{32}^{XZ}\gamma_{22}^{YZ}$ | $\delta_2^4\gamma_{32}^{XZ}\gamma_{32}^{YZ}$ | $\delta_2^6\gamma_{32}^{XZ}\gamma_{42}^{YZ}$ |
| $x_4$ | $\gamma_{42}^{XZ}$ | $\delta_2^3\gamma_{42}^{XZ}\gamma_{22}^{YZ}$ | $\delta_2^6\gamma_{42}^{XZ}\gamma_{32}^{YZ}$ | $\delta_2^9\gamma_{42}^{XZ}\gamma_{42}^{YZ}$ |
| $Z = z_3$ | | | | |
| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
| $x_1$ | 1 | $\gamma_{23}^{YZ}$ | $\gamma_{33}^{YZ}$ | $\gamma_{43}^{YZ}$ |
| $x_2$ | $\gamma_{23}^{XZ}$ | $\delta_3\gamma_{23}^{XZ}\gamma_{23}^{YZ}$ | $\delta_3^2\gamma_{23}^{XZ}\gamma_{33}^{YZ}$ | $\delta_3^3\gamma_{23}^{XZ}\gamma_{43}^{YZ}$ |
| $x_3$ | $\gamma_{33}^{XZ}$ | $\delta_3^2\gamma_{33}^{XZ}\gamma_{23}^{YZ}$ | $\delta_3^4\gamma_{33}^{XZ}\gamma_{33}^{YZ}$ | $\delta_3^6\gamma_{33}^{XZ}\gamma_{43}^{YZ}$ |
| $x_4$ | $\gamma_{43}^{XZ}$ | $\delta_3^3\gamma_{43}^{XZ}\gamma_{23}^{YZ}$ | $\delta_3^6\gamma_{43}^{XZ}\gamma_{33}^{YZ}$ | $\delta_3^9\gamma_{43}^{XZ}\gamma_{43}^{YZ}$ |

gdzie $x_a$ oraz $x_b$ stanowią kategorie odniesienia dla zmiennych $X$ oraz $Y$. Model ten posiada $t$ parametrów opisujących interakcję zmiennych $X$ oraz $Y$. Liczba stopni swobody dla tego modelu wynosi:

$$df = t(r-1)(c-1) - t.$$

Choć w centrum zainteresowania są zmienne $X$ oraz $Y$, warto prześledzić, jakie konsekwencje ma powyższe założenie w odniesieniu do związku dotyczącego pozostałych par dwóch zmiennych. Na przykład warunkowy stosunek szans $\Theta_{2(3)2}^{X(Y)Z}$. Okazuje się, że:

$$\Theta_{2(3)2}^{X(Y)Z} = \frac{\delta_2^2\gamma_{22}^{XZ}\gamma_{32}^{YZ} \cdot \delta_3^4\gamma_{33}^{XZ}\gamma_{33}^{YZ}}{\delta_2^4\gamma_{32}^{XZ}\gamma_{32}^{YZ} \cdot \delta_3^2\gamma_{23}^{XZ}\gamma_{33}^{YZ}} = \left(\frac{\delta_3}{\delta_2}\right)^2 \cdot \frac{\gamma_{22}^{XZ} \cdot \gamma_{33}^{XZ}}{\gamma_{32}^{XZ} \cdot \gamma_{23}^{XZ}}.$$

Bardziej ogólnie daje się pokazać, że:

$$\Theta_{i(j)k}^{X(Y)Z} = \left(\frac{\delta_{k+1}}{\delta_k}\right)^{(j-1)} \cdot \delta_{ik}^{XZ}. \tag{2.80}$$

Analogicznie warunkowy stosunek szans dla zmiennych $Y$ oraz $Z$ wynosi:

$$\Theta_{(i)jk}^{(X)YZ} = \left(\frac{\delta_{k+1}}{\delta_k}\right)^{(i-1)} \cdot \delta_{jk}^{YZ}. \tag{2.81}$$

Jak widać interakcja drugiego rzędu pomiędzy $X$ oraz $Z$ zależy od wartości zmiennej $Y$ a interakcja drugiego rzędu pomiędzy $Y$ oraz $Z$ zależy od wartości zmiennej $X$. Wyrażając powyższą hipotezę za pomocą lokalnego stosunku szans trzech zmiennych otrzymujemy:

$$\Theta_{ijk}^{XYZ} = \frac{\delta_{k+1}}{\delta_k}. \tag{2.82}$$

Zauważmy, że gdyby powyższy iloraz był stały dla wszystkich sąsiednich kategorii zmiennej $Z$ mielibyśmy do czynienia z modelem $[XYZ_U]$ wcześniej prezentowanym (porównaj 2.70).

W powyższym modelu związek pomiędzy zmiennymi $X$ oraz $Y$ w podzbiorowościach wyróżnionych względem zmiennej $Z$ jest opisywany przez prosty model jednakowej interakcji. Związek ten może być oczywiście opisany za pomocą innych, prezentowanych wcześniej hipotez, tj. efektu wierszowego (R), modelu kolumnowego (C), modelu wierszowo–kolumnowego pierwszego bądź drugiego typu (RC1 bądź RC2), bądź nieokreślonego wzoru zależności (FA).

Interesujące są możliwości formułowania interakcji trzeciego rzędu dla hipotezy o efekcie wierszowym (kolumnowym). Przypomnijmy, że zgodnie z formułą 2.26 lokalny stosunek szans w modelu wierszowym można przedstawić jako:

$$\Theta_{ij}^{XY} = \delta \cdot \delta_{i\cdot}$$

Parametr $\delta$ opisuje ogólną siłę zależności[20] związanej z poszczególnymi efektami wierszowymi. Parametry $\delta_{i\cdot}$ opisują relacje pomiędzy poszczególnymi efektami wierszowymi. Ten typ parametryzacji, pozwoli na sformułowanie interakcji trzeciego rzędu związanej z efektem wierszowym. Zacznijmy od sytuacji najprostszej: zależność pomiędzy zmiennymi $X$ oraz $Y$ jest opisywana przez efekt wierszowy i nie zależała od wartości trzeciej zmiennej. Wówczas:

$$\Theta_{ij(k)}^{XY(Z)} = \delta_{\cdots}^{XY(Z)} \cdot \delta_{i\cdot\cdot}^{XY(Z)} \tag{2.83}$$

Model taki opisywaliśmy omawiając kwestię modelowania interakcji drugiego rzędu. Jego parametryzację można sformułować następująco:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \xi^{(i-1)(j-1)} \cdot \phi_i^{(j-1)}, \tag{2.84}$$

[20]Przypomnijmy, że jest to nazwa umowna, co było sygnalizowane przy okazji omawiania modelu wierszowego.

przyjmując $x_1$, $y_1$, $z_1$ jako kategorie odniesienia. Warunkowa interakcja pomiędzy zmiennymi $X$ oraz $Y$ jest określona przez $r-1$ parametry. W odniesieniu do równania 2.84 można założyć $\phi_1 = \phi_r = 1$, wówczas parametry $\xi$ opisują średnią geometryczną lokalnych stosunków szans w każdej podzbiorowości wyróżnionej ze względu na zmienną $Z$[21]. Liczba stopni swobody omawianego modelu wynosi:

$$df = t(r-1)(c-1) - (r-1)$$

W literaturze nazywa się go często *homogenicznym modelem wierszowym* (Clogg 1982*b*), będzie on oznaczany jako $XY_{R1}|Z$ . Określenie „homogeniczny" dotyczy braku interakcji trzeciego rzędu. Nieco bardziej złożoną zależność opisuje *prosty heterogeniczny model wierszowy*, który oznaczać będziemy $XY_{R1}|Zh$. Zgodnie z nim:

$$\Theta_{ij(k)}^{XY(Z)} = \delta_{..k}^{XY(Z)} \cdot \delta_{i..}^{XY(Z)} \tag{2.85}$$

Parametr $\delta_{..k}^{XY(Z)}$ jest indeksowany przez wartości zmiennej $Z$, tak więc ogólna siła zależności związana z efektami wierszowymi zależy od podzbiorowości, w której badamy zależność. W konsekwencji parametry wierszowe są inne dla dowolnej wartości zmiennej $Z$. Okazuje się jednak, że iloraz dwóch lokalnych stosunków zmiennych $X$ i $Y$ nie zależy od tego, dla której wartości zmiennej $Z$ został wyróżniony, tj.

$$\frac{\Theta_{aj(k)}^{XY(Z)}}{\Theta_{bm(k)}^{XY(Z)}} = \frac{\delta_{..k}^{XY(Z)} \cdot \delta_{a..}^{XY(Z)}}{\delta_{..k}^{XY(Z)} \cdot \delta_{b..}^{XY(Z)}} = \frac{\delta_{a..}^{XY(Z)}}{\delta_{b..}^{XY(Z)}}, \tag{2.86}$$

dla każdej wartości $z_k$ i dowolnych wartości $x_a$, $x_b$, $y_j$, $y_m$. Należy zwrócić uwagę, że w powyższej formule, niekoniecznie musimy porównywać stosunki szans dla tej samej kategorii zmiennej $Y$, tj. nie musi zachodzić $y_j = y_m$. Oczywiście, tak samo jak w przypadku modelu wierszowego dla dwóch zmiennych, tak samo w przypadku warunkowej zależności lokalne stosunki szans wyodrębnione dla dwóch ustalonych wierszy $a$ oraz $a+1$ w tej samej podzbiorowości są takie same, tj. $\Theta_{ac(k)}^{XY(Z)} = \Theta_{ad(k)}^{XY(Z)}$. Formułę dotyczącą prawdopodobieństwa rozkładu łącznego w takim modelu można przedstawić jako:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \xi_k^{(i-1)(j-1)} \cdot \phi_i^{(j-1)}, \tag{2.87}$$

Warunkową interakcję pomiędzy zmiennymi $X$ oraz $Y$ w prostym heterogenicznym modelu wierszowym opisują $r-2$ parametry $\phi_i$ oraz $t$ parametry $\xi_k$. Przypomnijmy, ze po przyjęciu założenia $\phi_1 = \phi_r = 1$ parametry $\xi_k$ opisują średnią geometryczną lokalnych stosunków szans w poszczególnych podzbiorowościach zgodnych z rozkładem oczekiwanym. Liczba stopni swobody tego modelu wynosi:

$$df = t(r-1)(c-1) - (r-2) - t.$$

[21] Porównaj omówienie formuły 2.27.

W podobny sposób sformułować można *heterogeniczny model wierszowy* $(XY_{Rh1}|Zh)$, który głosi, że dla każdej wartości zmiennej $Z$ zależność pomiędzy $X$ oraz $Y$ daje się opisać, za pomocą efektu wierszowego, natomiast wielkość efektów wierszowych i relacje pomiędzy nimi mogą być różne dla kolejnych wartości $Z$, tj:

$$\Theta_{ij(k)}^{XY(Z)} = \delta_{..k}^{XY(Z)} \cdot \delta_{i \cdot k}^{XY(Z)} \tag{2.88}$$

Parametryzacja tej hipotezy może wyglądać następująco:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \phi_{ik}^{(j-1)}, \tag{2.89}$$

W odniesieniu do powyższej parametryzacji warunkową interakcję pomiędzy zmiennymi $X$ oraz $Y$ opisują $t(r-1)$ parametry $\phi_{ik}$. Wyrażając to w sposób analogiczny do formuł 2.84 oraz 2.87 mamy $t$ parametrów $\xi_k$, oraz $t(r-2)$ parametry $\phi_{ik}$, po przyjęciu założenia $\phi_{1k} = \phi_{rk} = 1$. Liczba stopni swobody tego modelu wynosi:

$$df = t(r-1)(c-1) - t(r-1)$$

Podsumowując: formułując hipotezę o efekcie wierszowym dla opisu związku pomiędzy $X$ oraz $Y$ musimy zdecydować, w jakim stopniu trzecia zmienna $Z$ różnicuje ten związek. Jeśli nie różnicuje, mamy do czynienia z *homogenicznym modelem wierszowym*, jeśli różnicuje tylko ogólną zależność natomiast parametry opisujące relacje między wierszami pozostają takie same — z *prostym modelem heterogenicznym*. Jeśli poszczególne parametry wierszowe są inne dla poszczególnych wartości trzeciej zmiennej wówczas model taki nazywany jest *heterogenicznym*. W podobny sposób można modelować interakcję trzeciego rzędu dotyczącą efektu kolumnowego oraz wierszowo-kolumnowego I i II typu. Tabela 2.26 stanowi zestawienie modeli, które daje się sformułować przy danym typie zależności, formułę lokalnego stosunku szans i liczbę stopni swobody dla każdego z tych modeli.

Hipoteza dla efektu wierszowo-kolumnowego I typu może dotyczyć tego, czy ogólna siła zależności jest taka sama ($\delta_{...}$) czy też różni się ($\delta_{..k}$) dla poszczególnych wartości trzeciej zmiennej. Oddzielnie można zakładać, że efekty kolumnowe i wierszowe są heterogeniczne bądź homogeniczne. W prostym heterogenicznym modelu wierszowo-kolumnowym $XY_{RCh1}|Z$ zakładamy, że jedynie ogólna siła związku jest heterogeniczna, tj:

$$\Theta_{ij(k)}^{XY(Z)} = \delta_{..k}^{XY(Z)} \cdot \delta_{i..}^{XY(Z)} \cdot \delta_{.j.}^{XY(Z)} \tag{2.90}$$

Analogicznie do 2.86 daje się pokazać, że iloraz dwóch lokalnych stosunków szans nie zależy od wartości trzeciej zmiennej :

Tabela 2.26: Modele opisujące warunkową zależność dwóch zmiennych $X$ i $Y$ ze względu na trzecią zmienną $Z$

| Modele logarytmiczno–liniowe | | | |
|---|---|---|---|
| Model | Oznaczenie modelu | $\Theta_{ij(k)}^{XY(Z)} = \ldots$ | Liczba stopni swobody |
| Homogeniczny jednakowej interakcji | $XY_{UA}\|Z$ | $\delta_{\cdot\cdot\cdot}^{XY(Z)}$ | $t(r-1)(c-1)-1$ |
| Heterogeniczny jednakowej interakcji | $XY_{UA}\|Zh$ | $\delta_{\cdot\cdot k}^{XY(Z)}$ | $t(r-1)(c-1)-t$ |
| Homogeniczny wierszowy | $XY_{R}\|Z$ | $\delta_{\cdot\cdot\cdot}^{XY(Z)} \cdot \delta_{i\cdot\cdot}^{XY(Z)}$ | $t(r-1)(c-1)-r-1$ |
| Prosty heterogeniczny wierszowy | $XY_{R}\|Zh$ | $\delta_{\cdot\cdot k}^{XY(Z)} \cdot \delta_{i\cdot\cdot}^{XY(Z)}$ | $t(r-1)(c-1)-(r-2)-t$ |
| Heterogeniczny wierszowy | $XY_{Rh}\|Zh$ | $\delta_{\cdot\cdot k}^{XY(Z)} \cdot \delta_{i\cdot k}^{XY(Z)}$ | $t(r-1)(c-1)-t(r-1)=t(r-1)(c-2)$ |
| Homogeniczny kolumnowy | $XY_{C}\|Z$ | $\delta_{\cdot\cdot\cdot}^{XY(Z)} \cdot \delta_{\cdot j\cdot}^{XY(Z)}$ | $t(r-1)(c-1)-c-1$ |
| Prosty heterogeniczny kolumnowy | $XY_{C}\|Zh$ | $\delta_{\cdot\cdot k}^{XY(Z)} \cdot \delta_{\cdot j\cdot}^{XY(Z)}$ | $t(r-1)(c-1)-(c-2)-t$ |
| Heterogeniczny kolumnowy | $XY_{Ch}\|Zh$ | $\delta_{\cdot\cdot k}^{XY(Z)} \cdot \delta_{\cdot jk}^{XY(Z)}$ | $t(r-1)(c-1)-t(c-1)=t(c-1)(r-2)$ |
| Homogeniczny wierszowo-kolumnowy | $XY_{RC1}\|Z$ | $\delta_{\cdot\cdot\cdot}^{XY(Z)} \cdot \delta_{i\cdot\cdot}^{XY(Z)} \cdot \delta_{\cdot j\cdot}^{XY(Z)}$ | $t(r-1)(c-1)-(r-2)-(c-2)-1$ |
| Prosty heterogeniczny wierszowo-kolumnowy | $XY_{RC1}\|Zh$ | $\delta_{\cdot\cdot k}^{XY(Z)} \cdot \delta_{i\cdot\cdot}^{XY(Z)} \cdot \delta_{\cdot j\cdot}^{XY(Z)}$ | $t(r-1)(c-1)-(r-2)-(c-2)-t$ |
| Heterogeniczne parametry wierszowe, homogeniczne kolumnowe | $XY_{RhC1}\|Zh$ | $\delta_{\cdot\cdot k}^{XY(Z)} \cdot \delta_{i\cdot k}^{XY(Z)} \cdot \delta_{\cdot j\cdot}^{XY(Z)}$ | $t(r-1)(c-1)-t(r-2)-(c-2)-t$ |
| Heterogeniczne parametry kolumnowe, homogeniczne wierszowe | $XY_{RCh1}\|Zh$ | $\delta_{\cdot\cdot k}^{XY(Z)} \cdot \delta_{i\cdot\cdot}^{XY(Z)} \cdot \delta_{\cdot jk}^{XY(Z)}$ | $t(r-1)(c-1)-(r-2)-t(c-2)-t$ |
| Heterogeniczny wierszowo-kolumnowy | $XY_{RhCh1}\|Zh$ | $\delta_{\cdot\cdot k}^{XY(Z)} \cdot \delta_{i\cdot k}^{XY(Z)} \cdot \delta_{\cdot jk}^{XY(Z)}$ | $t(r-2)(c-2)$ |
| Modele logarytmiczno–multiptikatywne | | | |
| Model | Oznaczenie modelu | $\Theta_{ij(k)}^{XY(Z)} = \ldots$ | Liczba stopni swobody |
| Homogeniczny wierszowo-kolumnowy | $XY_{RC2}\|Z$ | $\left(\delta^{XY}\right)^{(u_{i+1}-u_i)(v_{j+1}-v_j)}$ | $t(r-1)(c-1)-(r-2)-(c-2)-1$ |
| Prosty heterogeniczny wierszowy | $XY_{R2}\|Zh$ | $\left(\delta_k^{XY}\right)^{(u_{i+1}-u_i)v}$ | $t(r-1)(c-1)-(r-2)-t$ |
| Prosty heterogeniczny kolumnowy | $XY_{C2}\|Zh$ | $\left(\delta_k^{XY}\right)^{u(v_{j+1}-v_j)}$ | $t(r-1)(c-1)-(c-2)-t$ |
| Prosty heterogeniczny wierszowo-kolumnowy | $XY_{RC2}\|Zh$ | $\left(\delta_k^{XY}\right)^{(u_{i+1}-u_i)(v_{j+1}-v_j)}$ | $t(r-1)(c-1)-(r-2)-(c-2)-t$ |
| Heterogeniczne parametry wierszowe, homogeniczne kolumnowe | $XY_{RhC2}\|Zh$ | $\left(\delta_k^{XY}\right)^{(u_{(i+1)k}-u_{ik})(v_{j+1}-v_j)}$ | $t(r-1)(c-1)-t(r-2)-(c-2)-t$ |
| Heterogeniczne parametry kolumnowe, homogeniczne wierszowe | $XY_{RCh2}\|Zh$ | $\left(\delta_k^{XY}\right)^{(u_{i+1}-u_i)(v_{(j+1)k}-v_{jk})}$ | $t(r-1)(c-1)-(r-2)-t(c-2)-t$ |
| Heterogeniczny wierszowo-kolumnowy | $XY_{RhCh2}\|Zh$ | $\left(\delta_k^{XY}\right)^{(u_{(i+1)k}-u_{ik})(v_{(j+1)k}-v_{jk})}$ | $t(r-2)(c-2)$ |

$$\frac{\Theta^{XY(Z)}_{aj(k)}}{\Theta^{XY(Z)}_{bm(k)}} = \frac{\delta^{XY(Z)}_{..k} \cdot \delta^{XY(Z)}_{a\cdot\cdot} \cdot \delta^{XY(Z)}_{.j.}}{\delta^{XY(Z)}_{..k} \cdot \delta^{XY(Z)}_{b\cdot\cdot} \cdot \delta^{XY(Z)}_{.m.}} = \frac{\delta^{XY(Z)}_{a\cdot\cdot} \cdot \delta^{XY(Z)}_{.j.}}{\delta^{XY(Z)}_{b\cdot\cdot} \cdot \delta^{XY(Z)}_{.m.}}, \tag{2.91}$$

dla każdej wartości $z_k$ i dowolnych wartości $x_a$, $x_b$, $y_j$, $y_m$. Oczywiście inaczej niż w modelu wierszowym (kolumnowym) lokalne stosunki szans wyodrębnione dla dwóch dowolnych wierszy (kolumn) nie muszą być takie same. Kolejny, bardziej złożony model $XY_{RCh1}|Zh$ zakłada, że poza ogólną siłą zależności również parametry kolumnowe różnią się dla wartości trzeciej zmiennej, tj.

$$\Theta^{XY(Z)}_{ij(k)} = \delta^{XY(Z)}_{..k} \cdot \delta^{XY(Z)}_{i\cdot\cdot} \cdot \delta^{XY(Z)}_{.jk} \tag{2.92}$$

Iloraz dwóch lokalnych stosunków dotyczący tej samej wartości zmiennej kolumnowej nie zależy od wartości trzeciej zmiennej.

$$\frac{\Theta^{XY(Z)}_{aj(k)}}{\Theta^{XY(Z)}_{bj(k)}} = \frac{\delta^{XY(Z)}_{..k} \cdot \delta^{XY(Z)}_{a\cdot\cdot} \cdot \delta^{XY(Z)}_{.jk}}{\delta^{XY(Z)}_{..k} \cdot \delta^{XY(Z)}_{b\cdot\cdot} \cdot \delta^{XY(Z)}_{.jk}} = \frac{\delta^{XY(Z)}_{a\cdot\cdot}}{\delta^{XY(Z)}_{b\cdot\cdot}} \tag{2.93}$$

dla każdej wartości $z_k$ i dowolnych wartości $x_a$, $x_b$, $y_j$. Zauważmy, że w odróżnieniu od omawianych powyżej modeli — prostego heterogenicznego modelu wierszowego i wierszowo–kolumnowego tj. formuł 2.86 oraz 2.91 — iloraz w formule 2.93 dotyczy stosunków szans wyróżnionych dla tej samej wartości zmiennej kolumnowej. Inny model $XY_{RhC1}|Zh$ może uwzględniać heterogeniczne parametry zmiennej wierszowej, przy homogenicznych parametrach kolumnowych. Najbardziej złożony model logarytmiczno–liniowy $XY_{RhCh1}|Zh$ zamieszczony w tabeli 2.26 różnicuje zarówno parametry zmiennej wierszowej jak i kolumnowej.

Podobnie jak dla modeli logarytmiczno–liniowych, tak dla modeli logarytmiczno–multiplikatywnych kolejne hipotezy dotyczą heterogenicznej bądź homogenicznej siły związku i parametrów skalowania kategorii obydwu zmiennych. Warto zwrócić uwagę na trzy ostatnie modele przedstawione w tabeli 2.26 dopuszczające odmienne skalowanie wartości zmiennych $X$ oraz $Y$ dla różnych wartości trzeciej zmiennej. W heterogenicznym modelu wierszowo–kolumnowym formuła dotycząca lokalnego stosunku szans przedstawia się jako:

$$\Theta^{XY(Z)}_{ij(k)} = \delta_k^{(u_{(i+1)k}-u_{ik})(v_{(j+1)k}-v_{jk})} \tag{2.94}$$

Przypuśćmy, że zmienną, której wartości skalujemy jest wykształcenie a zmienną grupująca jest kraj bądź rok badania. Można oczekiwać, że odległości pomiędzy poszczególnymi kategoriami wykształcenia mogą się różnić pomiędzy krajami, ze względu na specyfikę systemów kształcenia w poszczególnych krajach. Podobnie można oczekiwać, że wiele czynników — np. reforma oświaty — może wpływać na to, że

odległości te zmieniają się w czasie. Przyjęcie heterogenicznego skalowania może być w takich sytuacjach uzasadnione.

Krótkiej wzmianki wymagają jeszcze modele $XY_{C2}|Zh$ oraz $XY_{R2}|Zh$. Omawiając modele dla dwóch zmiennych zaznaczyliśmy, że nałożenie na logarytmiczno–multiplikatywny model warunku o identycznych odległościach pomiędzy kolejnymi kategoriami zmiennej wierszowej, tj. $u_{i+1} - u_i = u$ prowadzi do sformułowania logarytmiczno-liniowego o efekcie kolumnowym (porównaj 2.55), analogicznie założenie nałożone na parametry skalujące dla zmiennej kolumnowej $v_{j+1} - v_j = v$ jest tożsame z hipotezą o efekcie wierszowym. Można zadać pytanie, czy analogicznie jest w przypadku modelowanie zależności warunkowych, a dokładniej:

1. Czy nałożenie na logarytmiczno–multiplikatywny model heterogeniczny wierszowo–kolumnowy $XY_{RhCh2}|Zh$ założenia $u_{(i+1)k} - u_{ik} = u_k$ prowadzi do sformułowania logarytmiczno–liniowego modelu heterogenicznego kolumnowego modelu $XY_{Ch}|Zh$?

2. Czy nałożenie na model logarytmiczno–multiplikatywny model $XY_{RCh2}|Zh$ założenia $u_{i+1} - u_i = u$ prowadzi do sformułowania logarytmiczno–liniowego modelu $XY_{Ch}|Zh$?

3. Czy nałożenie na model logarytmiczno–multiplikatywny model $XY_{RhC2}|Zh$ założenia $u_{(i+1)k} - u_{ik} = u_k$ prowadzi do sformułowania logarytmiczno–liniowego modelu $XY_{C}|Zh$?

4. Czy nałożenie na model logarytmiczno–multiplikatywny model *prosty* heterogeniczny wierszowo–kolumnowy $XY_{RC2}|Zh$ założenia $u_{i+1} - u_i = u$ prowadzi do sformułowania logarytmiczno–liniowego prostego kolumnowego modelu $XY_{C}|Zh$?

Analogicznie można sformułować cztery pytania dotyczące założenia o jednakowych odległościach pomiędzy kolejnymi kategoriami zmiennej kolumnowej $v_{j+1} - v_j = v$. Odpowiedź na dwa pierwsze z postawionych pytań jest pozytywna. Model $XY_{Ch1}|Zh$ zakłada jedynie, że w każdej podzbiorowości wyróżnionej ze względu na zmienną $Z$, lokalne stosunki szans dla dwóch ustalonych kolumn i dowolnych sąsiednich wierszy są takie same (tj. $\Theta^{XY(Z)}_{aj(k)} = \Theta^{XY(Z)}_{bj(k)}$), nie zakładamy natomiast niczego odnośnie relacji pomiędzy tymi podzbiorowościami. Zauważmy, że jeśli założymy, że $u_{(i+1)k} - u_{ik} = u_k$ w odniesieniu do modelu $XY_{RhCh2}|Zh$, w którym wszystkie parametry są heterogeniczne (indeksowane przez wartości trzeciej zmiennej) wówczas:

$$\frac{\Theta^{XY(Z)}_{aj(k)}}{\Theta^{XY(Z)}_{bj(k)}} = \frac{\delta_k^{(u_{(a+1)k}-u_{ak})(v_{(j+1)k}-v_{jk})}}{\delta_k^{(u_{(b+1)k}-u_{bk})(v_{(j+1)k}-v_{jk})}} = \frac{\delta_k^{u_k(v_{(j+1)k}-v_{jk})}}{\delta_k^{u_k(v_{(j+1)k}-v_{jk})}} = 1. \tag{2.95}$$

Tak więc zachodzi opisany powyżej warunek dotyczący heterogenicznego modelu kolumnowego. Podobnie jest on spełniony jeśli założenie $u_{i+1} - u_i = u$ nałożymy na model $XY_{RCh2}|Zh$, który głosi, że parametry wierszowe pozostają homogeniczne. Wówczas:

$$\frac{\Theta_{aj(k)}^{XY(Z)}}{\Theta_{bj(k)}^{XY(Z)}} = \frac{\delta_k^{(u_{(a+1)}-u_a)(v_{(j+1)k}-v_{jk})}}{\delta_k^{(u_{(b+1)}-u_b)(v_{(j+1)k}-v_{jk})}} = \frac{\delta_k^{u(v_{(j+1)k}-v_{jk})}}{\delta_k^{u(v_{(j+1)k}-v_{jk})}} = 1 \tag{2.96}$$

Odnosząc się do pytania trzeciego i czwartego należy przypomnieć, że model $XY_C|Zh$ w porównaniu do modelu $XY_{Ch}|Zh$, zakłada dodatkowo, że iloraz dwóch dowolnych stosunków szans dotyczących $X$ i $Y$ w każdej podzbiorowości wyróżnionej ze względu na zmienna $Z$ jest stały (porównaj 2.86). Okazuje się, że pomimo nałożenia warunku$u_{(i+1)k} - u_{ik} = u_k$ na model $XY_{RhC2}|Zh$ iloraz pomiędzy warunkowymi lokalnymi stosunkami szans zależy od rozpatrywanej podzbiorowości:

$$\frac{\Theta_{aj(k)}^{XY(Z)}}{\Theta_{bm(k)}^{XY(Z)}} = \frac{\delta_k^{(u_{(a+1)k}-u_{ak})(v_{(j+1)}-v_j)}}{\delta_k^{(u_{(b+1)k}-u_{bk})(v_{(m+1)}-v_m)}} = \frac{\delta_k^{u_k(v_{(j+1)}-v_j)}}{\delta_k^{u_k(v_{(m+1)}-v_m)}} \tag{2.97}$$

Podobnie jeśli nałożymy ten warunek na model $XY_{RC2|Zh}$ otrzymujemy,

$$\frac{\Theta_{aj(k)}^{XY(Z)}}{\Theta_{bm(k)}^{XY(Z)}} = \frac{\delta_k^{(u_{(a+1)}-u_a)(v_{(j+1)}-v_j)}}{\delta_k^{(u_{(b+1)}-u_b)(v_{(m+1)}-v_m)}} = \frac{\delta_k^{u(v_{(j+1)}-v_j)}}{\delta_k^{u(v_{(m+1)}-v_m)}} \tag{2.98}$$

Wynika z tego, że odpowiedź na zadane powyżej pytania trzecie i czwarte jest negatywna. W związku z tym konieczne jest wyodrębnienie hipotezy będącej połączeniem modelu $XY_{RhC2}|Zh$ i warunku $u_{i+1} - u_i = u$. W tabeli 2.26 została ona oznaczona jako model $XY_{C2}|Zh$. Połączeniem modelu $XY_{RC2}|Zh$ i warunku $u_{i+1} - u_i = u$ prowadzi do sformułowania identycznego modelu, należy bowiem zauważyć, że nie ma znaczenia czy parametr interakcji podnosimy do potęgi $u$ czy też $u_k$ (porównaj formuły 2.97 oraz 2.98), gdyż parametry skalujące są znormalizowane (porównaj warunki 2.47), nie zmienia to więc ogólności modelu. Analogicznie, połączenie modelu $XY_{RCh2}|Zh$ (lub $XY_{RC2}|Zh$) i warunku $v_{j+1} - v_j = v$ zostało wyszczególnione jako model $XY_{R2}|Zh$.

Kolejny przykład empiryczny zilustruje możliwości modelowania warunkowych zależności i interpretacji poszczególnych modeli. W tabeli 2.27 przedstawiony jest rozkład łączny liczebności zainteresowania polityką (zmienna $Y$) i wykształcenia (zmienna $X$) w 10 wybranych krajach europejskich (zmienna $Z$). Dane pochodzą z pierwszej tury Europejskiego Sondażu Społecznego, realizowanego w 2002 roku[22]. Odnośnie zainteresowania polityką respondenci poproszeni byli o odpowiedź na pytanie *Jak by Pan(i) określił(a) swoje zainteresowanie polityką? Czy Pan(i)...* polityką: ...Do wyboru były cztery odpowiedzi. 1. *bardzo się interesuje*, 2. *dość się interesuje*, 3. *niezbyt*

[22]Realizacja w niektórych krajach przeciągnęła się do 2003 roku.

Tabela 2.27: Wykształcenie ($X$) a zainteresowanie polityką ($Y$) w wybranych krajach europejskich[a]

| Belgia | | | | | Finlandia | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $X\backslash Y$ | 1 | 2 | 3 | 4 | $X\backslash Y$ | 1 | 2 | 3 | 4 |
| 1 | 12,0 | 61,0 | 87,0 | 122,0 | 1 | 15,0 | 144,0 | 172,0 | 79,0 |
| 2 | 17,0 | 102,0 | 144,0 | 109,0 | 2 | 21,0 | 126,0 | 187,0 | 54,0 |
| 3 | 42,0 | 267,0 | 258,0 | 131,0 | 3 | 43,0 | 267,0 | 315,0 | 72,0 |
| 4 | 83,0 | 256,0 | 131,0 | 44,0 | 4 | 65,0 | 247,0 | 164,0 | 21,0 |
| Grecja | | | | | Holandia | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | $X\backslash Y$ | 1 | 2 | 3 | 4 |
| 1 | 72,0 | 171,7 | 350,9 | 396,0 | 1 | 15,7 | 91,6 | 85,2 | 48,0 |
| 2 | 31,1 | 102,3 | 167,9 | 183,0 | 2 | 57,8 | 372,4 | 259,7 | 93,1 |
| 3 | 47,6 | 117,4 | 192,7 | 175,5 | 3 | 72,0 | 399,3 | 177,8 | 35,8 |
| 4 | 87,9 | 174,2 | 170,0 | 115,7 | 4 | 154,8 | 393,4 | 87,7 | 13,2 |
| Irlandia | | | | | Niemcy | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | $X\backslash Y$ | 1 | 2 | 3 | 4 |
| 1 | 28,7 | 133,0 | 124,8 | 157,6 | 1 | 3,8 | 9,9 | 26,7 | 14,3 |
| 2 | 30,5 | 170,9 | 160,8 | 167,8 | 2 | 33,2 | 139,2 | 183,2 | 62,2 |
| 3 | 44,3 | 178,0 | 159,8 | 83,5 | 3 | 302,0 | 700,3 | 560,1 | 97,3 |
| 4 | 79,2 | 285,8 | 159,1 | 75,7 | 4 | 262,8 | 393,2 | 117,9 | 6,7 |
| Polska | | | | | Szwecja | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | $X\backslash Y$ | 1 | 2 | 3 | 4 |
| 1 | 13,4 | 131,2 | 216,6 | 174,9 | 1 | 38,0 | 207,0 | 228,0 | 92,0 |
| 2 | 30,2 | 209,1 | 298,8 | 124,9 | 2 | 41,0 | 155,0 | 153,0 | 39,0 |
| 3 | 26,2 | 227,5 | 238,9 | 41,7 | 3 | 46,0 | 201,0 | 148,0 | 35,0 |
| 4 | 32,5 | 170,1 | 138,2 | 15,6 | 4 | 117,0 | 342,0 | 135,0 | 17,0 |
| Węgry | | | | | Włochy | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | $X\backslash Y$ | 1 | 2 | 3 | 4 |
| 1 | 22,0 | 141,0 | 212,0 | 180,0 | 1 | 10,4 | 37,0 | 92,5 | 136,4 |
| 2 | 28,0 | 169,0 | 206,0 | 67,0 | 2 | 21,5 | 75,4 | 160,2 | 155,3 |
| 3 | 35,0 | 182,0 | 133,0 | 31,0 | 3 | 42,5 | 133,6 | 150,2 | 81,3 |
| 4 | 47,0 | 119,0 | 47,0 | 8,0 | 4 | 19,9 | 49,5 | 22,8 | 9,6 |

[a]Europejski Sondaż Społeczny, 2002-2003. Dane przeważone.

Tabela 2.28: Wyniki weryfikacji wybranych modeli dla danych z tabeli 2.27, ESS, 2001.

| Modele logarytmiczno—liniowe | | | | |
|---|---|---|---|---|
| Model | df | $\chi^2$ | $G^2$ | $\Delta$ |
| $[XZ][YZ]$ | 90 | 1991,2 (p<0,0001) | 1990,8 (p<0,0001) | 11,1 |
| $[XY][XZ][YZ]$ | 81 | 283,2 (p<0,0001) | 285,6 (p<0,0001) | 4,0 |
| $XY_{UA}\|Z$ | 89 | 427,6 (p<0,0001) | 421,4 (p<0,0001) | 5,1 |
| $XY_{UA}\|Zh$ | 80 | 203,1 (p<0,0001) | 197,2 (p<0,0001) | 3,5 |
| $XY_R\|Z$ | 87 | 356,8 (p<0,0001) | 351,5 (p<0,0001) | 4,5 |
| $XY_R\|Zh$ | 78 | 152,8 (p<0,0001) | 150,4 (p<0,0001) | 2,9 |
| $XY_{Rh}\|Zh$ | 60 | 126,2 (p<0,0001) | 123,9 (p<0,0001) | 2,7 |
| $XY_C\|Z$ | 87 | 413,3 (p<0,0001) | 411,7 (p<0,0001) | 5,0 |
| $XY_C\|Zh$ | 78 | 174,4 (p<0,0001) | 172,4 (p<0,0001) | 3,4 |
| $XY_{Ch}\|Zh$ | 60 | 138,6 (p<0,0001) | 136,7 (p<0,0001) | 3,0 |
| $XY_{RC1}\|Z$ | 85 | 322,1 (p<0,0001) | 318,6 (p<0,0001) | 4,3 |
| $XY_{RC1}\|Zh$ | 76 | 98,8 (p=0,0406) | 98,9 (p=0,0399) | 2,4 |
| $XY_{RhC1}\|Zh$ | 58 | 73,5 (p=0,0828) | 73,4 (p=0,0834) | 2,2 |
| $XY_{RCh1}\|Zh$ | 58 | 58,7 (p=0,4506) | 57,6 (p=0,4884) | 1,8 |
| $XY_{RhCh1}\|Zh$ | 40 | 41,6 (p=0,4020) | 41,0 (p=0,4281) | 1,5 |
| Modele logarytmiczno—multiplikatywne | | | | |
| Model | df | $\chi^2$ | $G^2$ | $\Delta$ |
| $XY_{RC2}\|Z$ | 85 | 335,5 (p<0,0001) | 331,6 (p<0,0001) | 4,4 |
| $XY_{R2}\|Zh$ | 78 | 160,7 (p<0,0001) | 161,8 (p<0,0001) | 3,1 |
| $XY_{C2}\|Zh$ | 78 | 172,4 (p<0,0001) | 171,1 (p<0,0001) | 3,4 |
| $XY_{RC2}\|Zh$ | 76 | 109,8 (p=0,0068) | 109,3 (p=0,0074) | 2,6 |
| $XY_{RhC2}\|Zh$ | 58 | 75,6 (p=0,0603) | 74,8 (p=0,0680) | 2,1 |
| $XY_{RCh2}\|Zh$ | 58 | 77,6 (p=0,0435) | 75,9 (p=0,0571) | 2,0 |
| $XY_{RhCh2}\|Zh$ | 40 | 49,5 (p=0,1451) | 48,7 (p=0,1626) | 1,6 |

*się interesuje* 4. *w ogóle się nie interesuje*[23]. Jeśli chodzi o wykształcenie wyróżnione zostały cztery kategorie w oparciu o międzynarodowy schemat klasyfikacji ISCED[24].

[23]Z analizy wykluczone zostały osoby, które odpowiedziały „Trudno powiedzieć" bądź odmówiły odpowiedzi na to pytanie.

[24]Schemat ten został zarekomendowany w latach 70-tych XX wieku przez UNESCO. W badaniu wykorzystano wersję z 1997 r, która wyróżnia siedem kategorii: 1. nieukończone podstawowe, 2. podstawowe, 3. nieukończone średnie (w tym zasadnicze zawodowe), 4. ukończone średnie, 5. ponadśrednie, nie będące dające się zaklasyfikować jako wykształcenie wyższe, 6. nieukończone wyższe, licencjat, 7. pełne wyższe. Ponieważ liczebności niektórych kategorii w wybranych krajach były

Można oczekiwać, że ze względu na odmienny kontekst społeczno–kulturowy związek pomiędzy zainteresowaniem polityką a wykształceniem może się różnić w poszczególnych krajach. Na początku przetestowaliśmy dwie hipotezy: warunkowej niezależności obydwu zmiennych oraz identyczności warunkowych stosunków szans w poszczególnych krajach[25]. Wyniki weryfikacji przedstawione w tabeli 2.28 pokazują, że pierwsza z nich powinna być odrzucona, przy konwencjonalnie przyjmowanych poziomach istotności, również posługiwanie się indeksem rozbieżności prowadzi do podobnych konkluzji. W przypadku drugiej hipotezy, tj modelu $[XY][XZ][YZ]$ wartości statystyk $G^2$ oraz $\chi^2$ są znacznie mniejsze, jednak nawet na bardzo małym poziomie 0,0001 model ten należy odrzucić. Co prawda, posługujemy się bardzo dużą próbą (ponad 20 tys. jednostek) a w takim przypadku — z powodów wymienionych w pierwszym rozdziale — wnioski o odrzuceniu hipotezy należy podejmować ostrożnie. Jednak wartość indeksu rozbieżności (niezgodnie z modelem jest zaklasyfikowanych ok. 4% próby) nie wskazuje, że jest to model akceptowalny[26].

Ponieważ nie ma zdecydowanych przesłanek przemawiających za zaakceptowaniem hipotezy o identyczności związku (a dokładniej równości warunkowych stosunków szans) pomiędzy wykształceniem a zainteresowaniem polityką w poszczególnych krajach, celowe wydaje się przetestowanie hipotez przedstawionych w tabeli 2.26. Modele homogeniczne — jednakowej interakcji, wierszowy, kolumnowy i wierszowo–kolumnowe — są szczególnym przypadkiem modelu $[XY][XZ][YZ]$ gdyż, zakładają one brak różnic w związku pomiędzy poszczególnymi krajami, przy czym modelują ten związek w prostszy sposób.

Pozostałe modele — heterogeniczne — nie mogą być porównane z modelem $[XY][XZ][YZ]$ pod względem prostoty. Należy przypomnieć, że heterogeniczny model jednakowej interakcji, zakłada, że związek między zmiennymi $X$ i $Y$ może być opisany za pomocą jednego parametru, który interpretujemy jako lokalny stosunek szans. Zgodnie z tym modelem wielkość tego parametru w poszczególnych krajach może się różnić. Parametry dla poszczególnych krajów zostały przedstawione w tabeli 2.29. Porównajmy ze sobą Belgię i Finlandię: w pierwszym z tych dwóch krajów, lokalny stosunek szans wynosi w przybliżeniu 0,675 co wskazuje, że np. proporcja

bardzo małe, dla celów analizy konieczne było połączenie pierwszej i drugiej kategorii jak również kategorii 5-7.

[25]Modele prostsze są słabo dopasowane. Warto jednak zaznaczyć, że w odniesieniu do analizowanych danych nie jest sensowne testowanie hipotezy o równomierności zmiennej $Z$. Zmienna ta zdaje jedynie sprawę z liczebności próby w poszczególnych krajach, dlatego w modelu powinno się uwzględnić parametr związany z tą zmienną, tj. $d_k^Z$.

[26]Do podobnych konkluzji prowadzi użycie omówionego w rozdziale pierwszym miernika BIC, który nie został zaprezentowaliśmy w tabeli 2.28. Zgodnie z wartościami tego miernika wiele innych modeli ma lepsze dopasowanie do danych.

Tabela 2.29: Parametry heterogeniczne z modeli omawianych w tekście dla danych z tabeli 2.27

| Model | $XY_{UA}\|Zh$ | $XY_{R1}\|Zh$ | $XY_{RC1}\|Zh$ | $XY_{RCh1}\|Zh$: $\psi_{\cdot 1k}^{XY(Z)}$=1 | | | $XY_{RhCh2}\|Zh$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parametr | $\delta_k$ | $\xi_{\cdot\cdot k}^{XY(Z)}$ | $\xi_{\cdot\cdot k}^{XY(Z)}$ | $\psi_{\cdot 2k}^{XY(Z)}$ | $\psi_{\cdot 3k}^{XY(Z)}$ | $\psi_{\cdot 4k}^{XY(Z)}$ | $\delta_k$ | $u_1$ | $u_2$ | $u_3$ | $u_2$ | $v_1$ | $v_2$ | $v_3$ | $v_2$ |
| Belgia | 0,675 | 0,677 | 0,697 | 0,791 | 0,553 | 0,338 | 6,696 | 0,585 | 0,267 | -0,091 | -0,761 | -0,615 | -0,272 | 0,166 | 0,721 |
| Finlandia | 0,768 | 0,764 | 0,770 | 0,758 | 0,617 | 0,410 | 4,678 | 0,487 | 0,300 | 0,032 | -0,819 | -0,620 | -0,238 | 0,121 | 0,738 |
| Grecja | 0,847 | 0,838 | 0,852 | 0,925 | 0,734 | 0,614 | 0,401 | 0,460 | 0,316 | 0,053 | -0,828 | -0,589 | -0,346 | 0,248 | 0,687 |
| Holandia | 0,631 | 0,635 | 0,630 | 0,697 | 0,434 | 0,261 | 9,780 | 0,573 | 0,289 | -0,102 | -0,760 | -0,639 | -0,260 | 0,205 | 0,694 |
| Irlandia | 0,793 | 0,794 | 0,811 | 0,902 | 0,762 | 0,521 | 3,125 | 0,526 | 0,404 | -0,212 | -0,718 | -0,552 | -0,291 | 0,064 | 0,779 |
| Niemcy | 0,530 | 0,554 | 0,549 | 0,710 | 0,356 | 0,176 | 20,436 | 0,584 | 0,284 | -0,116 | -0,752 | -0,581 | -0,320 | 0,174 | 0,728 |
| Polska | 0,691 | 0,682 | 0,692 | 0,866 | 0,673 | 0,306 | 6,657 | 0,656 | 0,258 | -0,249 | -0,664 | -0,489 | -0,308 | -0,020 | 0,816 |
| Szwecja | 0,742 | 0,739 | 0,731 | 0,840 | 0,597 | 0,410 | 4,823 | 0,563 | 0,222 | 0,011 | -0,796 | -0,582 | -0,320 | 0,175 | 0,727 |
| Węgry | 0,635 | 0,620 | 0,618 | 0,734 | 0,471 | 0,226 | 11,403 | 0,642 | 0,211 | -0,127 | -0,726 | -0,589 | -0,278 | 0,117 | 0,750 |
| Włochy | 0,643 | 0,627 | 0,638 | 0,884 | 0,501 | 0,292 | 8,665 | 0,575 | 0,324 | -0,166 | -0,733 | -0,527 | -0,392 | 0,191 | 0,729 |

osób które *bardzo interesują się polityką* do osób które *dość się interesują* jest prawie 1,5 (tj. 1/0,675) razy większa wśród osób z wykształceniem niepełnym średnim w porównaniu do osób z wykształceniem podstawowym. W Finlandii siła tego związku jest mniejsza, tj. lokalny stosunek szans wynosi 0,768 (tj. jest bardziej zbliżona do wartości 1 czyli stanu niezależności stochastycznej).

Heterogeniczny model jednakowej interakcji jest jednak słabo dopasowany do danych. Mierniki przedstawione w tabeli 2.28, wskazują, że wzór zależności pomiędzy zmiennymi jest bardziej złożony. Możliwe jest uwzględnienie specyfiki zmiennej wierszowej lub kolumnowej. Mierniki dopasowania do danych dla heterogenicznych modeli wierszowych są generalnie lepsze aniżeli dla modeli kolumnowych. Wskazuje na to zarówno redukcja wielkości statystyki $G^2$ w stosunku do modelu jednakowej interakcji, jak również wielkości indeksów rozbieżności.

W prostym heterogenicznym modelu wierszowym, uwzględniona jest specyfika kategorii wykształcenia. Zgodnie z przyjętym kodowaniem parametry wierszowe dla dwóch skrajnych kategorii tj. wykształcenia podstawowego ($\phi_1^X$) i wyższego niż średnie ($\phi_4^X$) są równe 1, natomiast dla wykształcenia średniego $\phi_2^X = 1,104$, a dla wykształcenia niepełnego średniego $\phi_3^X = 1,147$. Jak pokazuje formuła 2.85 w modelu wierszowym prostym heterogenicznym odpowiednie parametry nie różnią się ze względu na trzecią zmienną (kraj), natomiast zmienna ta różnicuje ogólną siłę zależności. Opisują to parametry $\xi_k$ przedstawione w tabeli 2.29. Dzięki przyjętej parametryzacji wartości te możemy interpretować jako średnią lokalnych stosunków szans w danym kraju. Parametry te pozwalają zrekonstruować odpowiednie stosunki szans. Na przykład lokalny stosunek szans wyróżniony dla Belgii ($Z = 1$) porównujący dwie pierwsze kategorie zmiennej $X$ i dowolne dwie sąsiednie kategorie zmiennej $Y$ wynosi:

$$\Theta_{1j(1)}^{XY(Z)} = \frac{\phi_2^X}{\phi_1^X} \cdot \xi_{\cdot\cdot 1}^{XY(Z)} = \frac{1,104}{1} \cdot 0,677 = 0,748,$$

co oznacza, że zgodnie z tym modelem w Belgii wśród osób z wykształceniem średnim proporcja osób, które *bardzo się interesują polityką* do osób, które polityką *dość się interesują* jest prawie 1,33 (tj. 1/0,748) razy większa niż wśród osób z wykształceniem podstawowym. Podobny jest związek dla innych sąsiednich par kategorii opisującej zainteresowanie polityką. Porównując osoby z wykształceniem ukończonym średnim z osobami z wykształceniem niepełnym średnim analogiczne stosunki szans wynoszą $\Theta_{3/2;j(1)}^{XY(Z)} = 0,704$. Lokalne stosunki szans przedstawione zostały w tabeli 2.30.

Wartości parametrów wierszowych, pokazują, że im większa jest różnica w wykształceniu, tym większych różnic można się spodziewać jeśli chodzi o zainteresowanie polityką, np.

$$\Theta_{3/1;j(1)}^{X \quad Y(Z)} = \Theta_{1j(1)}^{XY(Z)} \cdot \Theta_{2j(1)}^{XY(Z)} = 0,748 \cdot 0,704 = 0,526.$$

Podobnie jest w innych krajach, jednak wartości stosunków szans będą inne. Na przykład, dla Finlandii analogiczne stosunki szans wynoszą:

$$\Theta_{1j(2)}^{XY(Z)} = \frac{\phi_2^X}{\phi_1^X} \cdot \xi_{\cdot\cdot 2}^{XY(Z)} = \frac{1,104}{1} \cdot 0,764 = 0,844,$$

oraz $\Theta_{3/2;j(k)}^{XY(Z)} = 0,794$. Ponieważ, parametry wierszowe są takie same dla wszystkich krajów — zgodnie z formułą 2.86 — relacje pomiędzy dowolnymi stosunkami szans nie zależą od tego, który kraj rozpatrujemy, np.

$$\frac{\Theta_{23(1)}^{XY(Z)}/\Theta_{11(1)}^{XY(Z)}}{\Theta_{23(2)}^{XY(Z)}/\Theta_{11(2)}^{XY(Z)}} = \frac{0,704/0,748}{0,794/0,844} = 1.$$

Bardziej złożony heterogeniczny model wierszowy, wprowadza 18 dodatkowych parametrów w porównaniu do prostego modelu heterogenicznego. Test warunkowy porównujący te modele pokazuje, że na poziomie istotności 0,05 nie ma konieczności wprowadzania parametrów wierszowych specyficznych dla każdego kraju: $G^2 = 150,4 - 123,9 = 26,4$ przy 18 stopniach swobody $p = 0,09$.

Spośród modeli wierszowo–kolumnowych na uwagę zasługuje model prosty heterogeniczny. Model ten można porównać za pomocą testu warunkowego z omówionym powyżej prostym heterogenicznym modelem wierszowym. Okazuje się, że dodatkowe uwzględnienie parametrów kolumnowych daje znaczącą redukcję statystyki $G^2 = 150,4 - 98,9 = 55,1$ przy 2 stopniach swobody, $p < 0,0001$. Przypomnijmy, że w modelu tym występują parametry wierszowe i kolumnowe, których wartość jest taka sama dla poszczególnych krajów. Dla kolejnych kategorii wykształcenia wartości te wynoszą: $\phi_1^X = 1$, $\phi_2^X = 1,135$, $\phi_3^X = 1,138$, $\phi_4^X = 1$; jeśli chodzi o zainteresowanie polityką parametry wynoszą $\psi_1^Y = 1$ dla kategorii *bardzo się interesuje*, $\psi_2^Y = 1,142$ dla kategorii *dość się interesuję* , $\psi_3^Y = 1,198$, dla osób które zadeklarowały, że *niezbyt się interesują* oraz $\psi_4^Y = 1$ dla kategorii *w ogóle się nie interesuję*. Ogólna siła związku, w tym modelu jest różna dla poszczególnych krajów i odpowiednie parametry dla omawianego modelu $XY_{RC1}|Zh$ zostały przedstawione w tabeli. 2.29. Podobnie jak wcześniej, dzięki przyjętej parametryzacji (założeniu $\phi_1^X = \phi_4^X = 1$, $\psi_1^X = \psi_4^X = 1$) parametry ogólnej siły związku można interpretować jako średnią lokalnych stosunków szans w poszczególnych krajach.

Tak jak poprzednio, możliwe jest zrekonstruowanie odpowiednich stosunków szans. Wielkość ta dla Belgii dla pierwszych dwóch kategorii obydwu zmiennych wynosi:

$$\Theta_{11(k)}^{XY(Z)} = \frac{\phi_2^X \psi_2^Y}{\phi_1^X \psi_1^Y} \xi_{\cdot\cdot 1}^{XY(Z)} = 0,903.$$

Tabela 2.30: Warunkowe stosunki szans $\Theta_{ij(k)}^{XY(Z)}$ zgodne z wybranymi modelami dla danych z tabeli 2.27 dla Belgii i Finlandii

| Heterogeniczny model jednakowej interakcji | | | | | | | |
|---|---|---|---|---|---|---|---|
| Belgia | | | | Finlandia | | | |
| $i\backslash j$ | 1 | 2 | 3 | $i\backslash j$ | 1 | 2 | 3 |
| 1 | 0,675 | 0,675 | 0,675 | 1 | 0,768 | 0,768 | 0,768 |
| 2 | 0,675 | 0,675 | 0,675 | 2 | 0,768 | 0,768 | 0,768 |
| 3 | 0,675 | 0,675 | 0,675 | 3 | 0,768 | 0,768 | 0,768 |
| Prosty heterogeniczny model wierszowy | | | | | | | |
| Belgia | | | | Finlandia | | | |
| $i\backslash j$ | 1 | 2 | 3 | $i\backslash j$ | 1 | 2 | 3 |
| 1 | 0,748 | 0,748 | 0,748 | 1 | 0,844 | 0,844 | 0,844 |
| 2 | 0,704 | 0,704 | 0,704 | 2 | 0,794 | 0,794 | 0,794 |
| 3 | 0,590 | 0,590 | 0,590 | 3 | 0,666 | 0,666 | 0,666 |
| Prosty heterogeniczny model wierszowo–kolumnowy | | | | | | | |
| Belgia | | | | Finlandia | | | |
| $i\backslash j$ | 1 | 2 | 3 | $i\backslash j$ | 1 | 2 | 3 |
| 1 | 0,903 | 0,797 | 0,699 | 1 | 0,998 | 0,881 | 0,772 |
| 2 | 0,830 | 0,733 | 0,642 | 2 | 0,917 | 0,810 | 0,710 |
| 3 | 0,660 | 0,583 | 0,511 | 3 | 0,730 | 0,644 | 0,564 |
| Heterogeniczne parametry kolumnowe, homogeniczne wierszowe | | | | | | | |
| Belgia | | | | Finlandia | | | |
| $i\backslash j$ | 1 | 2 | 3 | $i\backslash j$ | 1 | 2 | 3 |
| 1 | 0,910 | 0,805 | 0,704 | 1 | 0,872 | 0,937 | 0,765 |
| 2 | 0,830 | 0,734 | 0,642 | 2 | 0,796 | 0,854 | 0,698 |
| 3 | 0,655 | 0,579 | 0,506 | 3 | 0,627 | 0,673 | 0,550 |

W tabeli 2.30 przedstawione są stosunki pozostałe stosunki szans dla Belgii i Finlandii. Proporcja liczby osób, które *bardzo interesują się polityką* do liczby osób, które deklarują, że *dość się polityką* jest ponad 1,10 (tj. 1/0,903) razy mniejsza wśród osób z wykształceniem podstawowym w porównaniu do osób z wykształceniem niepełnym średnim, 1,2 (tj. 1/0,83) razy mniejsza wśród osób z wykształceniem niepełnym w porównaniu do osób z wykształceniem ukończonym średnim, oraz 1,5 (tj. 1/0,66) razy mniejsza wśród osób z wykształceniem ukończonym średnim w porównaniu do osób z wykształceniem wyższym. Podobnie jak poprzednio najbardziej różnią się od siebie

osoby z wykształceniem podstawowym i wyższym. W porównaniu do omawianego poprzednio modelu wierszowego prostego heterogenicznego lokalne stosunki szans dla dwóch kategorii zmiennej wierszowej nie są identyczne, jednak — zgodnie z formułą 2.92 — iloraz dwóch dowolnych stosunków szans nie zależy od tego, dla którego kraju, był on wyznaczony np.

$$\frac{\Theta_{23(1)}^{XY(Z)}/\Theta_{11(1)}^{XY(Z)}}{\Theta_{23(2)}^{XY(Z)}/\Theta_{11(2)}^{XY(Z)}} = \frac{0,642/0,903}{0,710/0,998} = 1.$$

Warunkowy test pokazuje, że na poziomie istotności 0,01 należałoby wprowadzić do opisanego powyżej modelu heterogeniczne parametry kolumnowe[27], tj. $G^2 = 98,9 - 57,6 = 41,3$ przy 18 stopniach swobody, p=0,001. Heterogeniczne parametry kolumnowe pozwalają uwzględnić specyfikę dotyczącą kategorii zmiennej zainteresowanie polityką. Parametry wierszowe dla tego modelu wynoszą $\phi_{1\cdot\cdot} = \phi_{4\cdot\cdot} = 1$, $\phi_{2\cdot\cdot} = 1,151$, $\phi_{1\cdot\cdot} = 1,208$. Odpowiednie parametry kolumnowe zostały przedstawione w tabeli 2.30. Zgodnie z tym modelem stosunek szans dla Belgii dla pierwszych dwóch kategorii obydwu zmiennych wynosi:

$$\Theta_{11(1)}^{XY(Z)} = \frac{\phi_{2\cdot\cdot}^{XY(Z)} \cdot \psi_{\cdot 21}^{XY(Z)}}{\phi_{1\cdot\cdot}^{XY(Z)} \cdot \psi_{\cdot 11}^{XY(Z)}} = 1,151 \cdot 0,791 = 0,91.$$

Analogiczny stosunek szans dla Finlandii wynosi:

$$\Theta_{11(2)}^{XY(Z)} = \frac{\phi_{2\cdot\cdot}^{XY(Z)} \cdot \psi_{\cdot 22}^{XY(Z)}}{\phi_{1\cdot\cdot}^{XY(Z)} \cdot \psi_{\cdot 12}^{XY(Z)}} = 1,151 \cdot 0,758 = 0,872.$$

Pozostałe lokalne stosunki szans dla Belgii i Finlandii przedstawiono w tabeli 2.30. Inaczej niż w poprzednich modelach, wielkość ilorazu dwóch stosunków szans nie zależy od kraju, tylko wówczas gdy porównywane są te same kolumny, np.

$$\frac{\Theta_{23(1)}^{XY(Z)}/\Theta_{13(1)}^{XY(Z)}}{\Theta_{23(2)}^{XY(Z)}/\Theta_{13(2)}^{XY(Z)}} = \frac{0,642/0,704}{0,698/0,765} = 1.$$

Przedstawione w tabeli 2.28 modele logarytmiczno–multiplikatywne stanowią alternatywne podejście do modelowania związku pomiędzy zmiennymi. Poszczególnych logarytmiczno–liniowych modeli wierszowo–kolumnowych nie można porównywać z ich logarytmiczno–multiplikatywnymi odpowiednikami za pomocą statystyki $G^2$. Nie będziemy szczegółowo omawiać tych modeli. Zaznaczmy, że odmienne skalowanie kategorii zmiennej wierszowej i kolumnowej w poszczególnych krajach może być uzasadnione. Tak jak sygnalizowaliśmy wcześniej, odległości pomiędzy kolejnymi kategoriami wykształcenia mogą się różnić, ze względu na specyfikę systemów kształcenia w

[27]Posługując się miernikiem BIC należałoby pozostać przy wyborze modelu prostego heterogenicznego modelu wierszowo–kolumnowego.

poszczególnych krajach. Co prawda zastosowany schemat kodowania ISCED został stworzony z myślą o prowadzeniu między–krajowych analiz porównawczych, to w praktyce trudno oczekiwać, aby kategorie te ściśle sobie odpowiadały. W odniesieniu do drugiej zmiennej skalowanie może użyteczne ze względu na różnice językowe w tłumaczeniu ankiety w poszczególnych krajach. Można przypuszczać, że sformułowania, które respondenci mają do wyboru aby opisać zainteresowanie polityką nie są ściśle porównywalne. W tabeli 2.29 przedstawione są wartości parametrów heterogenicznego modelu wierszowo–kolumnowego skalujących kategorie obydwu zmiennych i parametry opisujące siłę związku w poszczególnych krajach. Warto zauważyć, że przypisane parametry są monotoniczne względem kategorii dwóch zmiennych porządkowych.

### Propozycje uogólnienia modelowania warunkowej zależności

Kwestia modelowania warunkowej zależności była niejednokrotnie poruszana w literaturze. Najciekawsze propozycje uogólnienia pewnej klasy modeli tego typu przedstawili Yamaguchi (1987), Xie (1992), oraz Goodman i Hout (1998, 1998b, 2001). Autorzy ci, omawiają to zagadnienie w kontekście porównywania zjawiska ruchliwości społecznej w różnych krajach, bądź w różnych punktach czasowych, niemniej, modele przedstawione przez nich można również stosować do opisu innych zjawisk. Model Goodmana i Houta jest najbardziej ogólny w tym sensie że model Yamaguchiego oraz Xie są jego szczególnymi przypadkami. Choć model ten — jak zobaczymy poniżej — nie wyczerpuje możliwości modelowania warunkowej zależności, to obejmuje szereg modeli tego typu, nadając im interesującą interpretację graficzną.

Model Goodmana–Houta koncentruje się na możliwościach warunkowej modelowania interakcji pomiędzy dwiema zmiennymi. Zgodnie z tym modelem warunkowy stosunek szans zmiennych $X$ oraz $Y$ względem trzeciej zmiennej $Z$ — nazywanej w literaturze zmienną *warstwową* (*layer variable*) — można przedstawić jako:

$$\Theta_{ij(k)}^{XY(Z)} = \alpha_{ij} \cdot \left(\beta_{ij}^{XY}\right)^{\Upsilon_k} \tag{2.99}$$

dla dowolnych $j$, $j$, $k$.Ponieważ propozycja uogólnienia modelowania warunkowej zależności związana jest z pewnymi analogiami do regresji liniowej i jej interpretacjami graficznymi jej autorzy koncentrują się na formule dotyczącej logarytmu określonego powyżej warunkowego stosunku szans, tj.

$$\log \Theta_{ij(k)}^{XY(Z)} = a_{ij}^{XY} + b_{ij}^{XY} \cdot \Upsilon_k, \tag{2.100}$$

gdzie $a_{ij} = \log \alpha_{ij}$, $b_{ij} = \log \beta_{ij}$. W tym ujęciu widać analogie do modelowania liniowego: logarytm warunkowego stosunku szans jest liniową funkcją, względem wielkości $\Upsilon_k$, opisujących kolejne kategorie zmiennej warstwowej $Z$. Wielkości te mogą być

przypisane *a priori*, bądź są one szacowane w modelu. Aby lepiej zrozumieć specyfikę powyższych parametrów, zdefiniujmy za pomocą powyższych parametrów lokalny stosunek szans trzech zmiennych:

$$\Theta_{ijk}^{XYZ} = \frac{\Theta_{ij(k+1)}^{XY(Z)}}{\Theta_{ij(k)}^{XY(Z)}} = \frac{\alpha_{ij} \cdot \left(\beta_{ij}^{XY}\right)^{\Upsilon_{k+1}}}{\alpha_{ij} \cdot \left(\beta_{ij}^{XY}\right)^{\Upsilon_k}} = \left(\beta_{ij}^{XY}\right)^{\Upsilon_{k+1}-\Upsilon_k}, \tag{2.101}$$

Logarytm powyższego wyrażenia jest równy:

$$\log \Theta_{ijk}^{XYZ} = \log \Theta_{ij(k+1)}^{XY(Z)} - \log \Theta_{ij(k)}^{XY(Z)} = b_{ij}^{XY}(\Upsilon_{k+1} - \Upsilon_k). \tag{2.102}$$

Jak pokazują powyższe zapisy interakcję trzeciego rzędu definiują parametry „nachylenia" $b_{ij}^{XY}$ (odpowiednio $\beta_{ij}^{XY}$) oraz parametry zmiennej warstwowej $\Upsilon_k$. Parametry „przecięcia" $a_{ij}^{XY}$ nie mają wpływu na lokalny stosunek szans trzech zmiennych, opisują one interakcję drugiego rzędu.

Jak pokazali Goodman i Hout wcześniejsze propozycje Yamaguchiego (1987) i Xie (1992) są szczególnymi przypadkami zaproponowanego przez nich modelu. Porównując model Yamaguchiego do ogólnego modelu 2.100 w tym pierwszym zakłada się, że: $b_{ij} = const.$, tj:

$$\log \Theta_{ij(k)}^{XY(Z)} = a_{ij}^{XY} + \Upsilon_k, \tag{2.103}$$

lub równoważnie:

$$\Theta_{ij(k)}^{XY(Z)} = \alpha_{ij}^{XY} \cdot \eta_k \tag{2.104}$$

gdzie $\eta_k = exp(\Upsilon_k)$. W modelu zaproponowanym przez Xie zakłada się, że $a_{ij} = 0$, tj.

$$\log \Theta_{ij(k)}^{XY(Z)} = b_{ij}^{XY} \cdot \Upsilon_k, \tag{2.105}$$

lub równoważnie:

$$\Theta_{ij(k)}^{XY(Z)} = \left(\beta_{ij}^{XY}\right)^{\Upsilon_k}. \tag{2.106}$$

Porównując formuły 2.103 oraz 2.105 widzimy, że efekt zmiennej warstwowej jest bądź dodawany do parametrów opisujących interakcje dwóch zmiennych bądź przemnażany, stąd też modele Xie i Yamaguchiego nazywane są odpowiednio modelem *addytywnego* efektu warstwowego bądź modelem *multiplikatywnego* efektu warstwowego. Propozycję Yamaguchiego w najbardziej ogólnej formie można przedstawić w parametryzacji względem kategorii odniesienia jako:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \gamma_{ij}^{XY} \cdot \delta_k^{(i-a)(j-b)} \tag{2.107}$$

gdzie $x_a$ oraz $y_b$ są kategoriami odniesienia dla obydwu zmiennych. Powyższy zapis pokazuje, że w modelu tym zakładamy, że zmienne $X$ oraz $Y$ są zmiennymi porządkowymi. Analogicznie, parametryzacja najogólniejszej formy modelu zaproponowanego

przez Xie przedstawia się następująco:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \Psi_{ij}^{\Upsilon_k}, \tag{2.108}$$

gdzie parametr $\Psi_{ij}$ opisuje interakcję pomiędzy zmiennymi $X$ oraz $Y$. Model Goodmana–Houta można przedstawić jako:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \gamma_{ij}^{XY} \cdot \Psi_{ij}^{\Upsilon_k} \tag{2.109}$$

Goodman i Hout (1998) zwracają uwagę, że model Xie może być modelem nie–hierarchicznym. Porównanie modeli 2.108 oraz 2.109 pokazuje, że w modelu Xie nie występuje oddzielny parametr opisujący interakcję drugiego rzędu pomiędzy $X$ oraz $Y$, co może prowadzić do formułowania modeli trudnych w interpretacji. Model Yamaguchiego jest hierarchiczny, ale wymaga porządkowego pomiaru zmiennych $X$ oraz $Y$.

Dla lepszego zrozumienia specyfiki poszczególnych modeli porównane zostaną ilorazy logarytmów dwóch lokalnych stosunków szans trzech zmiennych wyznaczonych dla rożnych kombinacji zmiennych $X$ oraz $Y$ i tej samej kategorii zmiennej $Z$. Dla modelu Goodmana i Houta:

$$\frac{\log \Theta_{ijk}^{XYZ}}{\log \Theta_{mnk}^{XYZ}} = \frac{\log \Theta_{ij(k+1)}^{XY(Z)} - \log \Theta_{ij(k)}^{XY(Z)}}{\log \Theta_{mn(k+1)}^{XY(Z)} - \log \Theta_{mn(k)}^{XY(Z)}} = \frac{b_{ij}^{XY}(\Upsilon_{k+1} - \Upsilon_k)}{b_{mn}^{XY}(\Upsilon_{k+1} - \Upsilon_k)} = \frac{b_{ij}^{XY}}{b_{mn}^{XY}}, \tag{2.110}$$

Iloraz ten nie zależy od tego, z którą kategorią zmiennej warstwowej $Z$ mamy do czynienia. Podobny wniosek można sformułować dla modelu multiplikatywnego efektu, natomiast dla modelu efektu addytywnego powyższy iloraz jest zawsze równy 1, gdyż w modelu Yamaguchiego $b_{ij} = b_{mn}$.

Dla modelu efektu multiplikatywnego zaproponowanego przez Xie można natomiast zaobserwować, że iloraz dwóch logarytmów warunkowych stosunków szans wyznaczonych dla różnych kombinacji zmiennych $X$ oraz $Y$ i tej samej kategorii zmiennej $Z$, nie zależy od wartości trzeciej zmiennej, tj.

$$\frac{\log \Theta_{ij(k)}^{XY(Z)}}{\log \Theta_{mn(k)}^{XY(Z)}} = \frac{b_{ij}^{XY}\Upsilon_k}{b_{mn}^{XY}\Upsilon_k} = \frac{b_{ij}^{XY}}{b_{mn}^{XY}}, \tag{2.111}$$

Dla modelu addytywnego, bądź bardziej ogólnej propozycji Goodmana–Houta własność 2.111 nie musi być spełniona. Obydwa modele warstwowe — addytywny i multiplikatywny — można interpretować podobnie: wzór zależności pomiędzy $X$ i $Y$ jest taki sam dla każdej wartości zmiennej $Z$, choć siła tej zależności może być różna i określa ją parametr $\Upsilon_k$. Mówiąc dokładniej: w modelu Yamaguchiego, dla każdej podzbiorowości wyróżnionej ze względu na zmienną $Z$, iloraz dwóch dowolnych stosunków

szans dotyczący zmiennych $X$ i $Y$ jest taki sam, na co wskazuje formuła 2.104. W odniesieniu do modelu Xie, można sformułować podobny wniosek przy czym porównujemy iloraz dwóch *logarytmów* dowolnych stosunków szans, co pokazuje formuła 2.111.

Formuła 2.110 pozwala sprawdzić, czy dowolny model logarytmiczno–liniowy (bądź logarytmiczno–multiplikatywny) wchodzi w skład modeli, które można opisać za pomocą propozycji Goodmana–Houta. Co więcej, jeśli ten iloraz jest równy 1, mamy do czynienia z modelem Yamaguchiego, jeśli natomiast zachodzi zależność opisana w formule 2.111 mamy do czynienia z modelem Xie. Pod tym kątem można przeanalizować — przykładowo — omawiany wcześniej heterogeniczny model jednakowej interakcji (2.78). Zauważmy, że iloraz 2.110 jest równy 1, tak więc model ten może być postrzegany jako szczególny modelu Yamaguchiego. W parametryzacji 2.107 można przyjąć, że $\gamma_{ij}^{XY} = \delta$, czyli wielkość tę można potraktować jako wielkość interakcji pomiędzy zmiennymi $X$ oraz $Y$ dla kategorii odniesienia zmiennej warstwowej $Z$. Parametry $\delta_k$ wskazują na modyfikację interakcji dla kolejnych wartości tej zmiennej.

Heterogeniczny model jednakowej interakcji jest również szczególnym przypadkiem multiplikatywnego efektu warstwowego proponowanego przez Xie. W tym modelu iloraz 2.111 porównujący dwa logarytmy warunkowych stosunków szans jest równy:

$$\frac{\log \Theta_{ij(k)}^{XY(Z)}}{\log \Theta_{mn(k)}^{XY(Z)}} = \frac{\delta_k}{\delta_k} = 1,$$

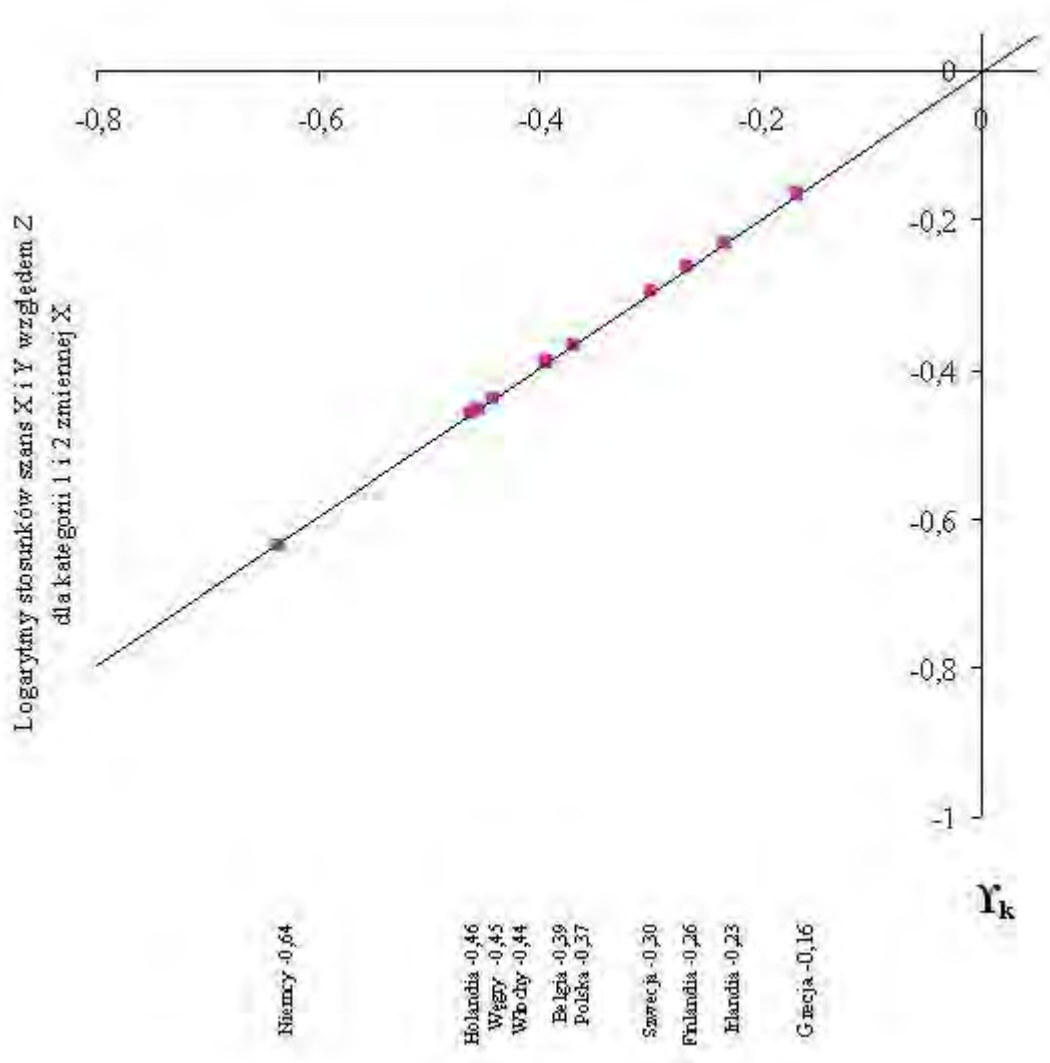
więc nie zależy od wartości trzeciej zmiennej. Zgodnie z konwencją zapisu modelu efektu multiplikatywnego zaproponowanego przez Xie model ten można przedstawić jako:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \left(\delta^{(i-a)(j-b)}\right)^{\Upsilon_k}$$

co jest równoważne parametryzacji 2.79. Model ten jest prostszy niż ogólny model multiplikatywnego efektu warstwowego 2.108, gdyż zachodzi $\psi_{ij} = \delta^{(i-a)(j-b)}$. Do ilustracji graficznej modelu heterogenicznej jednakowej interakcji — a ogólniej modelu Goodmana–Houta — wykorzystane zostaną dane z tabeli 2.27.

Na rysunku 2.6 przedstawione zostały logarytmy lokalnych stosunków szans zgodnych z heterogenicznym modelem jednakowej interakcji. Na osi poziomej zostały rozmieszczone kraje zgodnie z wartościami parametrów $\Upsilon_k$. W przypadku tego modelu parametry te są równe logarytmom parametrów $\delta_k$, czyli parametrom addytywnej postaci modelu logarytmiczno–liniowego. Wartości tych parametrów przedstawione są w tabeli 2.29, na przykład dla Belgii $\Upsilon_1 = ln(0,675) = -0,393$, itd. Warto zwrócić uwagę na kilka charakterystycznych cech tego wykresu: logarytmy warunkowych

Rysunek 2.6: Warunkowe logarytmy stosunków szans dla zmiennych $X$ oraz $Y$ względem $Z$ (heterogeniczny model jednakowej interakcji dane z tabeli 2.27)



stosunków szans leżą na linii prostej, co stanowi potwierdzenie, tego, ze jest to model Goodmana–Houta. Ponieważ model heterogenicznej jednakowej interakcji jest szczególnym przypadkiem modelu Yamaguchiego, parametr nachylenia linii prostej jest równy 1. Ponadto, ponieważ daje się go również interpretować jako szczególny przypadek modelu Xie daje się zauważyć, że parametr przecięcia jest równy 0. W odniesieniu do ogólnego modelu Goodmana–Houta:

$$\log \Theta_{ij(k)}^{XY(Z)} = a_{ij}^{XY} + b_{ij}^{XY} \cdot \Upsilon_k,$$

parametr $a_{ij}^{XY} = 0$ i $b_{ij}^{XY} = 1$, czyli $\Theta_{ij(k)}^{XY(Z)} = \Upsilon_k$. Ponieważ w modelu jednakowej interakcji wszystkie lokalne stosunki szans dla $X$ i $Y$ są sobie równe wykres ten jest adekwatny dla każdej pary kategorii obydwu zmiennych.

W heterogenicznym modelu jednakowej interakcji wartości $\Upsilon_k$ są szacowane w modelu. Jeśli zmienna $Z$ jest zmienną porządkową można dodatkowo założyć, że $\delta_k = \delta^k$, co prowadzi do sformułowania omawianego wcześniej modelu 2.70. Wielkości $\Upsilon_k$ byłyby wówczas równomiernie oddalone, tak jak na rysunku 2.5. Ponieważ model 2.70 jest osadzony w modelu heterogenicznej jednakowej interakcji model ten równiez może być interpretowany w kategoriach modeli Goodmana–Houta, Xie i Yamaguchiego.

Jak zostało pokazane, powyższe modele są zarówno modelami addytywnego jak też multiplikatywnego efektu wierszowego Pod tym kątem przeanalizowany zostanie teraz prosty heterogeniczny logarytmiczno–liniowy model wierszowy $XY_R|Zh$: Dla tej hipotezy iloraz 2.110 jest równy 1, tj.

$$\frac{\log \Theta_{ijk}^{XYZ}}{\log \Theta_{mnk}^{XYZ}} = \frac{\log \left[ \left( \xi_{..(k+1)}^{XY(Z)} \cdot \phi_{i\cdot\cdot}^{XY(Z)} \right) / \left( \xi_{..k}^{XY(Z)} \cdot \phi_{i\cdot\cdot}^{XY(Z)} \right) \right]}{\log \left[ \left( \xi_{..(k+1)}^{XY(Z)} \cdot \phi_{m\cdot\cdot}^{XY(Z)} \right) / \left( \xi_{..k}^{XY(Z)} \cdot \phi_{m\cdot\cdot}^{XY(Z)} \right) \right]} = 1$$

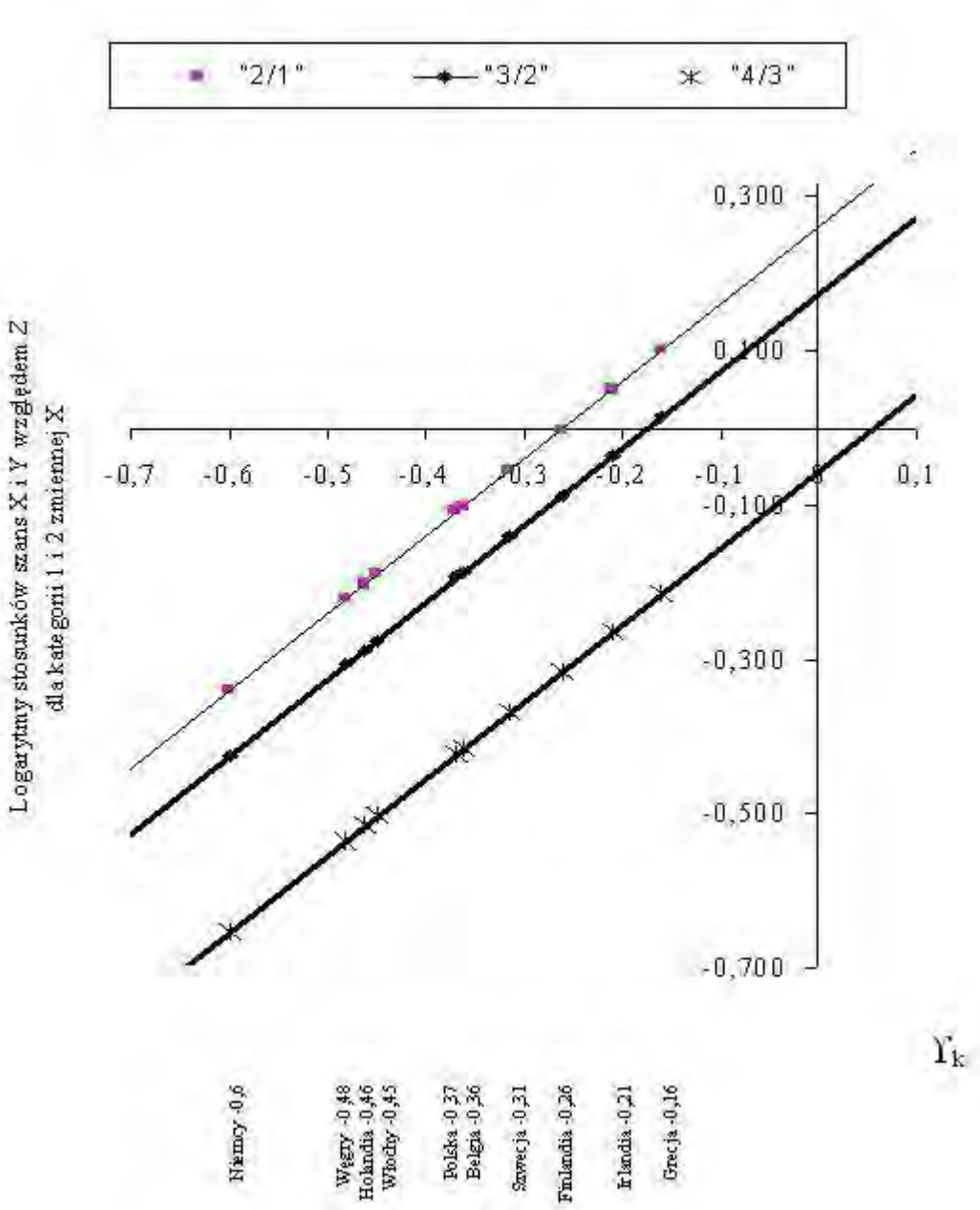
Natomiast iloraz 2.111

$$\frac{\log \Theta_{ij(k)}^{XY(Z)}}{\log \Theta_{mn(k)}^{XY(Z)}} = \frac{\log \left( \xi_{..k}^{XY(Z)} \cdot \phi_{i\cdot\cdot}^{XY(Z)} \right)}{\log \left( \xi_{..k}^{XY(Z)} \cdot \phi_{m\cdot\cdot}^{XY(Z)} \right)}$$

zależy od wartości zmiennej warstwowej $Z$. Prosty heterogeniczny model wierszowy może więc być postrzegany jako szczególny przypadek modelu Yamaguchiego, natomiast nie jest zgodny modelem zaproponowanym przez Xie. Podobna uwaga dotyczy prostych logarytmiczno–liniowych modeli kolumnowych i wierszowo–kolumnowych, gdyż w modelach tych jedynie parametr ogólnej siły związku pomiędzy $X$ i $Y$ jest indeksowany przez wartości zmiennej $Z$.

Rysunek 2.7 stanowi ilustrację dla prostego heterogenicznego modelu wierszowo–kolumnowego i danych z tabeli 2.27. Logarytmy lokalnych stosunków szans zgodnych z tym modelem leżą na jednej linii względem parametrów $\Upsilon_k$. Wartości $\Upsilon_k$ — zgodnie

Rysunek 2.7: Warunkowe logarytmy stosunków szans dla zmiennych $X$ oraz $Y$ względem $Z$ (prosty heterogeniczny model $XY_{RC1}|Zh$, dane z tabeli 2.27)



z którymi rozmieszczono kraje na osi poziomej — są parametrami addytywnego modelu i opisują ogólną siłę związku pomiędzy $X$ i $Y$. Są one logarytmami parametrów $\xi_{..k}^{XY(Z)}$ przedstawionych w tabeli 2.29. Na przykład dla Belgii ln(0,697)=-0,36, itd. Na wykresie na osi pionowej zamieszczono jedynie stosunki szans dla dwóch pierwszych kategorii zmiennej wierszowej $X$, tj. porównujące osoby z wykształceniem podstawowym i niepełnym średnim, oddzielnie dla wszystkich sąsiednich kategorii zmiennej $Y$, (pierwszej i drugiej, drugiej i trzeciej, trzeciej i czwartej) czyli wyrażenia $\Theta_{11(k)}^{XY(Z)}$, $\Theta_{12(k)}^{XY(Z)}$, $\Theta_{13(k)}^{XY(Z)}$. Zauważmy, że wykresy dla każdej pary zmiennej $Y$ są równoległe do siebie. Podobnie byłoby dla pozostałych lokalnych stosunków szans (wyznaczonych dla innych kategorii zmiennej $X$), które nie zostały zamieszczone na wykresie, co wynika z tego, że $b_{ij}^{XY} = 1$ i model ten można zaklasyfikować jako model Yamaguchiego. Parametr przecięcia w tym modelu definiują odpowiednie parametry wierszowe i kolumnowe:

$$a_{ij}^{XY} = \log\left(\frac{\phi_{(i+1)\cdot\cdot}^{XY(Z)} \cdot \psi_{\cdot(j+1)\cdot}^{XY(Z)}}{\phi_{i\cdot\cdot}^{XY(Z)} \cdot \psi_{\cdot j\cdot}^{XY(Z)}}\right) \tag{2.112}$$

Tak więc logarytm stosunku szans $\Theta_{11(1)}$ dla Belgii wynosi:

$$\begin{aligned} log\left(\Theta_{11(1)}^{XY(Z)}\right) &= \Upsilon_1 + a_{11}^{XY} = \log \xi_{\cdot\cdot(k)}^{XY(Z)} + \log\left(\frac{\phi_{2\cdot\cdot}^{XY(Z)} \cdot \psi_{\cdot 2\cdot}^{XY(Z)}}{\phi_{1\cdot\cdot}^{XY(Z)} \cdot \psi_{\cdot 1\cdot}^{XY(Z)}}\right) = \\ &= \log 0,697 + \log(1,135 \cdot 1,142) = -0,102. \end{aligned}$$

Modele heterogeniczne, w których nie tylko ogólna siła związku, ale też parametry wierszowe (kolumnowe), zależą od wartości zmiennej warstwowej $Z$, nie są modelami o strukturze zaproponowanej przez Goodmana–Houta. Np. dla heterogenicznego modelu wierszowego iloraz 2.110 jest równy:

$$\frac{\log \Theta_{ijk}^{XYZ}}{\log \Theta_{mnk}^{XYZ}} = \frac{\log\left[\left(\xi_{\cdot\cdot(k+1)}^{XY(Z)} \cdot \phi_{i\cdot(k+1)}^{XY(Z)}\right) / \left(\xi_{\cdot\cdot k}^{XY(Z)} \cdot \phi_{i\cdot k}^{XY(Z)}\right)\right]}{\log\left[\left(\xi_{\cdot\cdot(k+1)}^{XY(Z)} \cdot \phi_{m\cdot(k+1)}^{XY(Z)}\right) / \left(\xi_{\cdot\cdot k}^{XY(Z)} \cdot \phi_{m\cdot k}^{XY(Z)}\right)\right]}$$

Wielkość ta zależy od wartości zmiennej $Z$. W takiej sytuacji logarytmy stosunków szans dla danej kombinacji zmiennych $X$ oraz $Y$ niekoniecznie dają się przedstawić na linii prostej.

Przykładem modelu, który nie jest zgodny ze strukturą modelu addytywnego efektu warstwowego, a jest szczególnym przypadkiem modelu zaproponowanego przez Xie jest prosty heterogeniczny logarytmiczno–muliplikatywny model wierszowo-

kolumnowy. Iloraz 2.110 nie jest równy 1:

$$\begin{aligned}\frac{\log \Theta_{ijk}^{XYZ}}{\log \Theta_{mnk}^{XYZ}} &= \frac{\log\left[\left(\delta_{k+1}^{XY}\right)^{(u_{i+1}-u_i)(v_{j+1}-v_j)}\right] - \log\left[\left(\delta_{k}^{XY}\right)^{(u_{i+1}-u_i)(v_{j+1}-v_j)}\right]}{\log\left[\left(\delta_{k+1}^{XY}\right)^{(u_{m+1}-u_m)(v_{n+1}-v_n)}\right] - \log\left[\left(\delta_{k}^{XY}\right)^{(u_{m+1}-u_m)(v_{n+1}-v_n)}\right]} \\ &= \frac{(u_{i+1}-u_i)(v_{j+1}-v_j)\left[\delta_{(k+1)}^{XY} - \delta_{k}^{XY}\right]}{(u_{m+1}-u_m)(v_{n+1}-v_n)\left[\delta_{(k+1)}^{XY} - \delta_{k}^{XY}\right]} \\ &= \frac{(u_{i+1}-u_i)(v_{j+1}-v_j)}{(u_{m+1}-u_m)(v_{n+1}-v_n)} \qquad (2.113)\end{aligned}$$

Natomiast wartość ilorazu 2.111 nie zależy od wartości trzeciej zmiennej:

$$\frac{\log \Theta_{ij(k)}^{XY(Z)}}{\log \Theta_{mn(k)}^{XY(Z)}} = \frac{\log\left[\left(\delta_{k}^{XY}\right)^{(u_{i+1}-u_i)(v_{j+1}-v_j)}\right]}{\log\left[\left(\delta_{k}^{XY}\right)^{(u_{m+1}-u_m)(v_{n+1}-v_n)}\right]} = \frac{exp[(u_{i+1}-u_i)(v_{j+1}-v_j)]}{exp[(u_{m+1}-u_m)(v_{n+1}-v_n)]}$$

Na wykresie 2.8 przedstawiona została ilustracja tego modelu. Podobnie jak poprzednio wykorzystane zostały dane z tabeli 2.27. Kraje zostały rozmieszczone - analogicznie jak na poprzednich ilustracjach - na osi poziomej zgodnie z wartościami $\Upsilon_k$, w przypadku tego modelu $\Upsilon_k = log(\delta_k)$ (odpowiednie wartości parametrów można odczytać z wykresu). Podobnie jak na poprzednim wykresie przedstawione są stosunki szans $\Theta_{11(k)}^{XY(Z)}$, $\Theta_{12(k)}^{XY(Z)}$, $\Theta_{13(k)}^{XY(Z)}$, parametr przecięcia jest równy $b_{ij}^{XY} = 0$, podobnie byłoby dla każdego innego lokalnego stosunku szans nie przedstawionego na wykresie. Parametr nachylenia definiują w tym modelu odpowiednie parametry skalujące:

$$b_{ij}^{XY} = (u_{i+1} - u_i)(v_{j+1} - v_j) \qquad (2.114)$$

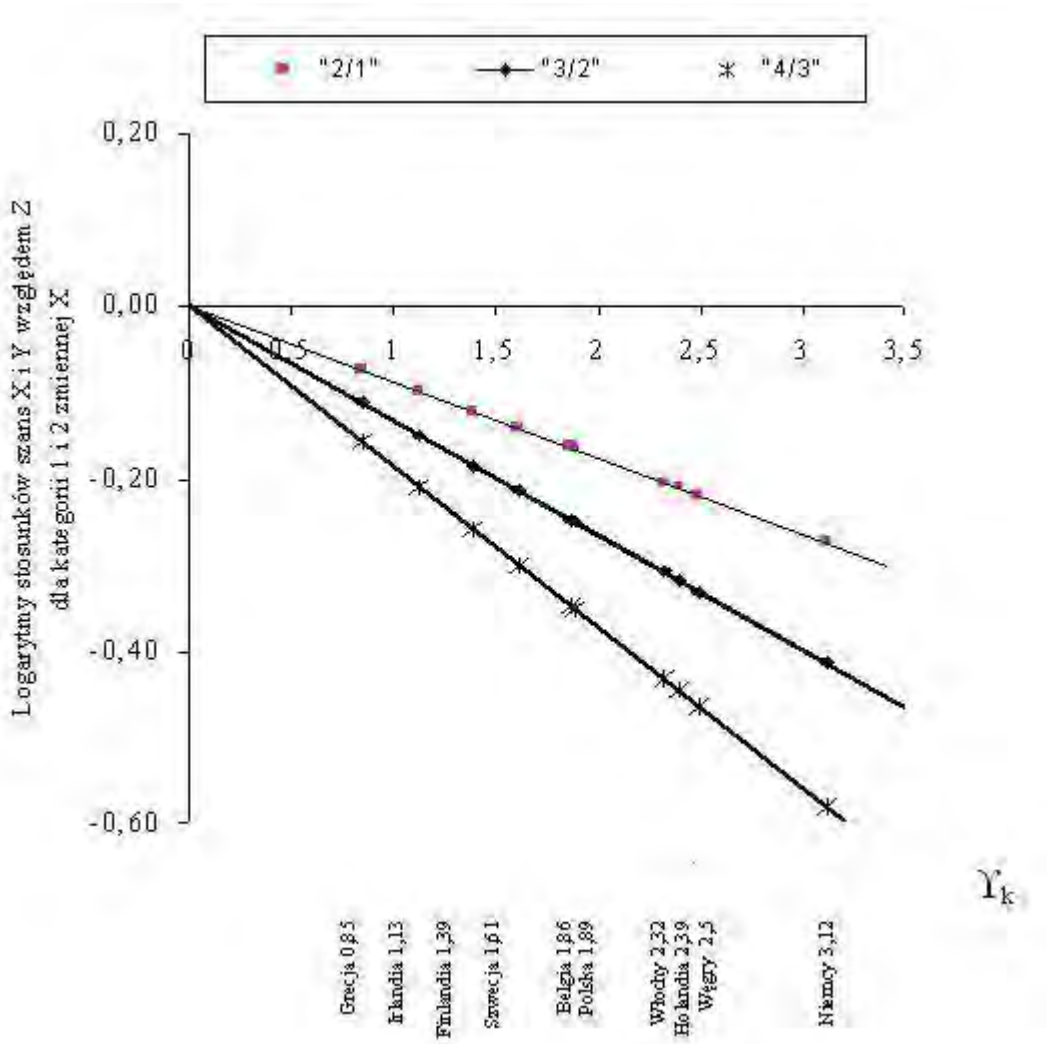
Parametry te nie były wcześniej przytoczone, wynoszą one $u_1 = 0,585$, $u_2 = 0,278$, $u_3 = -0,108$, $u_4 = -0,754$ dla zmiennej wierszowej i $v_1 = -0,584$, $v_2 = -0,296$, $v_3 = 0,137$, $v_4 = 0,744$. Przykładowo dla Belgii logarytm stosunku szans $\Theta_{11(1)}$ wynosi:

$$\begin{aligned}\log\left(\Theta_{11(1)}^{XY(Z)}\right) &= \Upsilon_1 \cdot b_{11}^{XY} = (\log \delta_k)(u_2 - u_1)(v_2 - v_1) = \qquad (2.115) \\ &= 1,86 \cdot (0,278 - 0,585)(-0,296 + 0,584) = -0,164.\end{aligned}$$

Równocześnie można pokazać, że logarytmiczno-multiplikatywny model, w którym parametry skalujące jednej bądź dwóch zmiennych $X$, $Y$ zależą od wartości zmiennej $Z$ nie dają się przedstawić jako model Goodmana–Houta.

Warto na koniec zauważyć, ze model Goodmana–Houta umożliwia formułowanie modeli, które nie są zgodne ani ze strukturą modelu addytywnego ani mulitplikatywnego efektu warstwowego. Autorzy wskazują (1998), że ich propozycja pozwala

Rysunek 2.8: Warunkowe logarytmy stosunków szans dla zmiennych $X$ oraz $Y$ względem $Z$ (prosty heterogeniczny logarytmiczno–multiplikatywny model wierszowo- kolumnowy, dane z tabeli 2.27)



formułować modele, w których interakcja drugiego rzędu i trzeciego rzędu może być opisywana za pomocą zależności innego typu. Zauważmy, ze nie pozwala na to model Xie, w którym interakcja drugiego i trzeciego rzędu nie jest modelowana oddzielnie. W propozycji Yamaguchiego, interakcja drugiego rzędu jest co prawda modelowana niezależnie od interakcji trzeciego rzędu, za to nie pozwala ona na modelowanie bardziej złożonej interakcji pomiędzy $X$ oraz $Y$. Model Goodmana–Houta nie ma takich ograniczeń. Autorzy podają przykład modelu, w którym interakcja drugiego rzędu jest opisywana za pomocą jednakowej interakcji (UA), a interakcja trzeciego rzędu ma nieokreślony wzór (FA). Zgodnie z tym modelem:

$$\Theta_{ij(k)}^{XY(Z)} = \alpha \cdot \left(\beta_{ij}^{XY}\right)^{\Upsilon_k} \tag{2.116}$$

W wersji addytywnej:

$$\log \Theta_{ij(k)}^{XY(Z)} = a^{XY} + b_{ij}^{XY} \cdot \Upsilon_k, \tag{2.117}$$

Jak widzimy parametr „przecięcia" w tym modelu nie jest specyficzny dla kombinacji obydwu zmiennych. Parametryzację tego modelu możemy przedstawić następująco:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \delta^{ij} \cdot \Psi_{ij}^{\Upsilon_k} \tag{2.118}$$

Propozycja Goodmana–Houta pozwala na uogólnienie i ciekawą interpretację szerokiej klasy modeli opisujących warunkową zależność. Model Yamaguchiego, a szczególnie model Xie są często wykorzystywane do modelowania interakcji trzeciego rzędu. Przykłady zastosowania modelu zaproponowanego przez Xie zostaną podane w kolejnym rozdziale tej pracy.

# Rozdział 3

# Modele dla tablic ruchliwości i danych panelowych

W poprzednich rozdziałach przedstawione zostały modele, pozwalające analizować rozkład łączny dwóch lub większej liczby zmiennych. Niektóre z nich były adekwatne w sytuacji, gdy wszystkie bądź wybrane zmienne mierzone były na skali porządkowej, inne nie czyniły takich założeń, tj. mogły być stosowane nawet wówczas, gdy wszystkie zmienne mierzone były na skali nominalnej. W modelach tych nie zakładało się nic więcej co do charakteru tych zmiennych.

W tym rozdziale przedstawione zostaną modele, które można formułować w odniesieniu do tablic ruchliwości i danych panelowych. Należy zauważyć, ze dane tego typu mają specyficzną strukturę: analizowane zmienne (zmienna wierszowa i kolumnowa) mają takie same kategorie. Na przykład, analizując międzypokoleniową ruchliwość zawodową, możliwe jest wyodrębnienie takich samych kategorii określających zawód ojca i zawód syna. W odniesieniu do danych o takiej strukturze interesująca może być na przykład odpowiedź na pytanie, jaki odsetek osób o pochodzeniu robotniczym wykonuje pracę tego typu. Podobną strukturę mają dane panelowe, gdyż tę samą cechę respondenta, można mierzyć w kilku punktach czasowych, na przykład można zestawić ze sobą odpowiedzi tych samych osób na to samo pytanie w kilku badaniach przeprowadzanych w określonych odstępach czasowych.

Istnieje wiele metod analizy danych tego typu. Modelowanie logarytmiczno–liniowe wydaje się analizą szczególnie przydatną, gdyż pozwala na sformułowanie wielu hipotez dotyczących rozkładu zmiennych i związku pomiędzy nimi. Z jednej strony, do analizy tablic ruchliwości i danych panelowych wykorzystywać można modele zaprezentowane do tej pory, z drugiej strony możliwe jest formułowanie nowych, specyficznych modeli, które nie byłyby adekwatne, gdyby kategorie zmiennych były odmienne.

Na początku tego rozdziału przedstawionych zostanie kilka przykładów tablic, w których kategorie zmiennych są takie same. Omówione zostaną podstawowe charakterystyki procentowe, za pomocą których można opisywać tego rodzaju tabele. Jednocześnie podkreślone zostaną ograniczenia wynikające z analizy „procentowej". Jak się okazuje, modelowanie logarytmiczno–liniowe w dużej mierze pozwala te ograniczenia wyeliminować. W pierwszej kolejności zaprezentowane zostaną modele jakie można sformułować w odniesieniu do rozkładu łącznego dwóch zmiennych. Choć niektóre z nich można zastosować do analizy zmiennych nominalnych, większość z nich wykorzystuje informację o porządkowym charakterze zmiennych.

Na końcu przedstawione zostaną możliwości rozszerzenia analizy na większą liczbę zmiennych. Z jednej strony trzecią zmienną może być zmienna grupująca, przykładowo możliwa jest analiza porównawcza ruchliwości w kilku podzbiorowościach, na przykład w kilku krajach. Z drugiej strony można analizować rozkład łączny trzech lub większej liczby zmiennych o takich samych kategoriach. Przykładowo dane badania panelowego przeprowadzonego trzy razy pozwalają na zestawienie ze sobą zmiennej mierzonej w trzech punktach czasowych. Ponadto, można porównywać zmiany w czasie dotyczące rozkładu nie jednej a kilku zmiennych. Na przykład, możliwe jest postawienie pytania, jak zmienił się związek pomiędzy wykształceniem i miejscem zamieszkania w dwóch punktach czasowych.

## 3.1 Przykłady tabel o takich samych kategoriach zmiennych

Prezentację rozpocznę podając kilka przykładów tablic, w których obydwie zmienne — wierszowa i kolumnowa — mają identyczne kategorie. Choć w dalsza część tego rozdziału poświęcona zostanie analizie tablic ruchliwości i danych panelowych to warto podkreślić, że wiele modeli i metod analizy zaprezentowanych w tym rozdziale ma zastosowanie w odniesieniu do danych dotyczących innych zjawisk, ale posiadających podobną strukturę. Na przykład, modele stosowane do badania ruchliwości społecznej wykorzystuje się do analizy wzorów zawierania małżeństw, gdzie kategorie opisujące wykształcenia męża korespondują z kategoriami wykształcenia żon.

Tabela 3.1 w sposób ogólny przedstawia strukturę danych, gdzie obydwie analizowane zmienne mają takie same kategorie. W konsekwencji liczba kategorii obydwu zmiennych jest identyczna i wynosi $r$. Warto zwrócić uwagę na komórki znajdujące się na przekątnej tej tabeli (zostały one pogrubione). Obejmują one obiekty, które posiadają takie same kategorie obydwu zmiennych. Odsetek takich obiektów w zbiorowości wynosi $\sum_{i=1}^{r} \pi_{ii}^{XY}$.

Dla pozostałych obiektów wartości obydwu analizowanych zmiennych różnią się. Jeśli analizowane zmienne mierzone są na skali porządkowej bądź mocniejszej sensowne staje się wyróżnienie obiektów, dla których jedna ze zmiennych ma wyższą (niższą) wartość od drugiej zmiennej. Zauważmy, że komórki powyżej tej przekątnej obejmują obiekty, dla których zmienna $Y$ przyjmuje wyższą wartość niż zmienna $X$, ich odsetek wynosi:

$$\sum_{i=1}^{r-1}\sum_{j>i}^{r}\pi_{ij}^{XY}.$$

Analogicznie, poniżej tej przekątnej znajdują się obiekty, dla których zmienna $X$ przyjmuje wyższą wartość niż zmienna $Y$, ich odsetek wynosi:

$$\sum_{j=1}^{r-1}\sum_{i>j}^{r}\pi_{ij}^{XY}.$$

Tabela 3.1: Rozkład łączny dwóch zmiennych $X$ i $Y$ o takich samych kategoriach

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $\ldots$ | $y_r$ | $\Sigma$ |
|---|---|---|---|---|---|---|
| $x_1$ | $\pi_{\mathbf{11}}^{\mathbf{XY}}$ | $\pi_{12}^{XY}$ | $\pi_{13}^{XY}$ | | $\pi_{1r}^{XY}$ | $\pi_1^X$ |
| $x_2$ | $\pi_{21}^{XY}$ | $\pi_{\mathbf{22}}^{\mathbf{XY}}$ | $\pi_{23}^{XY}$ | | $\pi_{2r}^{XY}$ | $\pi_2^X$ |
| $x_3$ | $\pi_{31}^{XY}$ | $\pi_{32}^{XY}$ | $\pi_{\mathbf{33}}^{\mathbf{XY}}$ | | $\pi_{3r}^{XY}$ | $\pi_3^X$ |
| $\vdots$ | | | | | | |
| $x_r$ | $\pi_{r1}^{XY}$ | $\pi_{r2}^{XY}$ | $\pi_{r3}^{XY}$ | | $\pi_{\mathbf{rr}}^{\mathbf{XY}}$ | $\pi_r^X$ |
| $\Sigma$ | $\pi_1^Y$ | $\pi_2^Y$ | $\pi_3^Y$ | | $\pi_r^Y$ | 1 |

Komórki powyżej i poniżej głównej przekątnej, można również scharakteryzować określając, o ile kategorii wartość zmiennej $X$ jest większa bądź mniejsza od wartości zmiennej $Y$. Na przykład prawdopodobieństwa $\pi_{12}^{XY}$, $\pi_{23}^{XY}$, $\ldots\pi_{(r-1)r}^{XY}$ wskazują na odsetek obiektów dla których zmienna $Y$ ma wartość wyższą o jedną kategorię od zmiennej $X$. Podobnie można wyznaczyć inne zbiory obiektów, np. prawdopodobieństwa $\pi_{(i+2)i}^{XY}$, tj. $\pi_{31}^{XY}$, $\pi_{42}^{XY}$, $\ldots$, itd. opisują obiekty dla których zmienna $X$ jest większa o dwie kategorie od zmiennej $Y$. Podzbiory komórek tego typu będą nazywane *pseudo–przekątnymi* [1]. Pojęcie to będzie użyteczne w dalszej części tego rozdziału, gdyż wiele modeli będzie odwoływało się do komórek określonych w ten sposób. Będziemy więc mówić o pseudo–przekątnych położonych bliżej lub dalej od

[1] W literaturze określa się je również jako „mniejsze przekątne" (*minor –diagonal*), co może być mylące, gdyż nie jest zgodne z rozumieniem przekątnej w geometrii

głównej przekątnej, na przykład prawdopodobieństwa $\pi^{XY}_{(i+1)i}$ opisują komórki poniżej głównej przekątnej przylegające do niej, a prawdopodobieństwa $\pi^{XY}_{i(i+3)}$ opisują pseudo–przekątną położoną o trzy komórki „w górę" od głównej przekątnej.

W dalszej części istotne będzie rozróżnienie lokalnych stosunków szans — czyli wyznaczonych dla sąsiednich kategorii jednej i drugiej zmiennej — ze względu na to, których komórek one dotyczą. Lokalne stosunki szans *wyznaczone* dla komórek na głównej przekątnej, to takie, dla których kategorie, które pojawiają się w indeksie dolnym zawierają tę sama wartość zmiennej wierszowej i kolumnowej, tj. $\Theta^{XY}_{ii}$, przykładowo $\Theta^{XY}_{11}$, $\Theta^{XY}_{22}$ itd. Podobnie można mówić o stosunkach szans wyznaczonych dla konkretnej pseudo–przekątnej. Przykładowo wielkości $\Theta^{XY}_{13}$, $\Theta^{XY}_{24}$ dotyczą stosunków szans, które w oznaczeniu mają komórki pseudo–przekątnej położonej o dwie komórki w górę od głównej przekątnej.

Zdefiniowane w ten sposób wielkości będą odróżniane od lokalnych stosunków szans *obejmujących* komórki przekątnej, czyli takich, które w swojej formule zawierają jedną bądź dwie komórki położone na przekątnej. Przykładowo stosunek szans $\Theta^{XY}_{12}$ nie jest wyznaczony dla komórki głównej przekątnej, ale obejmuje prawdopodobieństwo $\pi^{XY}_{22}$. Rozróżnienia te będą przydatne dla formułowania modeli w dalszej części tego rozdziału.

Tabela powyższa określa ogólną strukturę rozkładu łącznego gdy obydwie komórki mają te same kategorie, kolejne tabele będą stanowiły różne przykłady danych tego typu. Tabela 3.2 stanowi przykład danych opisujących ruchliwość społeczną. Na ogół wyróżnia się dwa typy ruchliwości:

1. między–pokoleniową — gdy porównuje się pozycję społeczną ojca (ewentualnie matki) i pozycję syna (ewentualnie córki)

2. wewnątrz–pokoleniową — gdy porównuje się pozycję społeczną badanej osoby w dwóch punktach czasowych.

W tej pracy będę się koncentrował na przykładach dotyczących ruchliwości między–pokoleniowej, można natomiast zauważyć, że ruchliwość wewnątrz–pokoleniowa może być traktowana jako szczególny przykład danych panelowych.

W kolejnych kolumnach tabeli 3.2 wymienione są kategorie opisujące przynależność społeczno–zawodową syna, w kolejnych wierszach — przynależność społeczno–zawodową ojca[2]. Tabelę tę, do której będę często odwoływał się w dalszej części tej

[2]Na ogół uwzględnia się pozycję zawodową ojca, gdy respondent miał 14 lat. Możliwe są oczywiście inne porównania np. pozycji córki i ojca, bądź córki i matki. Zwykło się porównywać pozycję ojca i syna, ponieważ wpływ pozycji ojca na pozycję społeczną kobiety jest mniej oczywisty niż w przypadku pozycji mężczyzny. Ponadto, istotne są względy praktyczne: wiele matek nigdy nie praco-

Tabela 3.2: Przynależność społeczno-zawodowa ojca i syna – tablica ruchliwości zawodowej[a]

| Przynależność społeczno–zawodowa ojca | Przynależność społeczno–zawodowa syna | | | | | | Suma |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| 1. Inteligencja i wyższe kadry kierownicze | **38** | 23 | 6 | 21 | 2 | 6 | 96 |
| 2. Pozostali pracownicy umysłowi | 48 | **65** | 13 | 80 | 6 | 10 | 222 |
| 3. Prywatni przedsiębiorcy | 10 | 6 | **13** | 25 | 3 | 5 | 62 |
| 4. Robotnicy wykwalifikowani | 35 | 85 | 29 | **294** | 26 | 20 | 489 |
| 5. Robotnicy niewykwalifikowani | 8 | 24 | 4 | 75 | **22** | 12 | 145 |
| 6. Właściciele gospodarstw i robotnicy rolni | 25 | 48 | 21 | 257 | 39 | **195** | 585 |
| Suma | 164 | 251 | 86 | 752 | 98 | 248 | 1599 |

[a]Źródło: Struktura społeczna II, 1987.

pracy będę nazywał w skrócie tablicą ruchliwości zawodowej. Dane pochodzą z badania „Struktura społeczna II" zrealizowanego w 1987 roku na ogólnopolskiej imiennej próbie 5884 jednostek (Słomczyński i inni, 1989 )[3]. Tabela 3.2 dotyczy mężczyzn w wieku 26-50 lat, gdyż można przypuszczać, że wzory ruchliwości mogą być związane z wiekiem, tj. mogą być różne dla przedstawicieli różnych kohort. Z tego względu często postuluje się, aby w jednej tabeli uwzględniać przedstawicieli możliwie jednorodnej grupy pod względem wieku, aby łatwiej było interpretować uzyskane wyniki. Kategorie społeczno–zawodowe wyodrębnione zostały zgodnie z klasyfikacją zawodów SKZ (Pohoski i inni 1974, Pohoski i Słomczyński 1978, Domański i Sawiński 1995, Domański i inni 2007).

Liczebności komórek na przekątnej odniesione do ogólnej liczebności próby, wskazują na tzw. odsetki dziedziczenia, tj. odsetek osób w próbie o danej przynależno-

wało co ograniczałoby analizowany zbiór danych. Wskazuje się również, że zawód mniej adekwatnie określa pozycję społeczną kobiety, gdyż jest mniej trwałym aspektem ich roli. Niemniej badania i analizy dotyczące ruchliwości kobiet były również przeprowadzane. Kwestia ruchliwości kobiet została omówiona m. in. w pracy Henryka Domańskiego (2007*b*). Wydaje się, że w przyszłości, wraz z zachodzącymi zmianami społecznymi zagadnienie to będzie zyskiwało na znaczeniu.

[3]W badaniu zastosowano próbę rezerwową, na ogół nie stosuje się juz zabiegu tego typu, ze względu na brak wystarczającego uzasadnienia teoretycznego.

ści społeczno–zawodowej, które odziedziczyły pozycję po swoim ojcu[4]. Przykładowo: 2,3% osób (tj. 38/1599) spośród przebadanych mężczyzn można zaklasyfikować, tak samo jak ich ojców do kategorii „inteligencja i wyższe warstwy". Odsetki dziedziczenia dla innych kategorii wynoszą odpowiednio: 4,1% dla pozostałych pracowników umysłowych, 0,8% dla prywatnych przedsiębiorców, 18,4% dla robotników wykwalifikowanych, 1,4% dla robotników niewykwalifikowanych oraz 12,2% dla rolników.

Łącznie, odsetek synów posiadających ten sam zawód co ich ojciec, tj. mężczyzn *niemobilnych* wynosi 39,2%. Wielkość ta stanowi pewną charakterystykę struktury społecznej. Niski odsetek osób niemobilnych zwykło się interpretować jako przejaw otwartości, gdyż wynik taki w pewnym zakresie potwierdzałby, że pochodzenie w małym stopniu determinuje pozycję syna. W omawianym przykładzie odsetek ten wynosi prawie 40%, czyli jest on znaczący biorąc pod uwagę, że oblicza się go sumując częstości jedynie 7 z 49 komórek rozkładu łącznego. Jak pokazują liczne badania dotyczące struktury społecznej (Hout 1983, Goodman 1972*c*, Domański 2007*b*), tendencja do dziedziczenia pozycji jest niemal uniwersalna.

Tabela 3.3: Odsetki napływu wyznaczone dla tablicy ruchliwości zawodowej (3.2)

| Przynależność społeczno–<br>–zawodowa ojca | Przynależność społeczno–<br>–zawodowa syna | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1. Inteligencja<br>i wyższe kadry kierownicze | 23,2 | 9,2 | 7,0 | 2,8 | 2,0 | 2,4 |
| 2. Pozostali pracownicy<br>umysłowi | 29,3 | 25,9 | 15,1 | 10,6 | 6,1 | 4,0 |
| 3. Prywatni przedsiębiorcy | 6,1 | 2,4 | 15,1 | 3,3 | 3,1 | 2,0 |
| 4. Robotnicy wykwalifikowani | 21,3 | 33,9 | 33,7 | 39,1 | 26,5 | 8,1 |
| 5. Robotnicy niewykwalifikowani | 4,9 | 9,6 | 4,7 | 10,0 | 22,4 | 4,8 |
| 6. Właściciele gospodarstw<br>i robotnicy rolni | 15,2 | 19,1 | 24,4 | 34,2 | 39,8 | 78,6 |
| Suma | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |

Innymi charakterystykami procentowymi opisującymi tablicę ruchliwości, są tzw. odsetki *napływu* i *odpływu*. Są to odpowiednio rozkłady warunkowe zawodu ojca

[4]Oczywiście nie chodzi o „dziedziczenie" w ścisłym znaczeniu tego słowa. Chodzi o sytuację, gdy syn należy do tej samej kategorii społeczno-zawodowej co jego ojciec, bez przesądzania, czy jest to efektem jakiegoś konkretnego procesu społecznego.

Tabela 3.4: Odsetki odpływu wyznaczone dla tablicy ruchliwości zawodowej (3.2)

| Przynależność społeczno–<br>–zawodowa ojca | Przynależność społeczno–<br>–zawodowa syna | | | | | | Suma |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| 1. Inteligencja<br>i wyższe kadry kierownicze | 39,6 | 24,0 | 6,3 | 21,9 | 2,1 | 6,3 | 100,0 |
| 2. Pozostali pracownicy<br>umysłowi | 21,6 | 29,3 | 5,9 | 36,0 | 2,7 | 4,5 | 100,0 |
| 3. Prywatni przedsiębiorcy | 16,1 | 9,7 | 21,0 | 40,3 | 4,8 | 8,1 | 100,0 |
| 4. Robotnicy wykwalifikowani | 7,2 | 17,4 | 5,9 | 60,1 | 5,3 | 4,1 | 100,0 |
| 5. Robotnicy niewykwalifikowani | 5,5 | 16,6 | 2,8 | 51,7 | 15,2 | 8,3 | 100,0 |
| 6. Właściciele gospodarstw<br>i robotnicy rolni | 4,3 | 8,2 | 3,6 | 43,9 | 6,7 | 33,3 | 100,0 |

względem zawodu syna i rozkłady warunkowe zawodu syna względem zawodu ojca. Prezentują je tabele 3.3 i 3.4. Przykładowo, odsetki napływu informują, że ojcowie 23,2% mężczyzn, których zaklasyfikować do grupy „inteligencja i wyższe warstwy", posiadali taką samą pozycję zawodową, ojcowie 6,1% byli prywatnymi przedsiębiorcami, itd. Natomiast odsetki odpływu wskazują, że spośród respondentów o pochodzeniu inteligenckim 39,6% odziedziczyło tę pozycję, a np. robotnicy niewykwalifikowani stanowią 2,1%.

Odsetki wyżej wymienione - dziedziczenia, odpływu i napływu - posiadają pewne wady, gdyby chciało się je interpretować jako mierniki ruchliwości. Po pierwsze, ich wielkości zależą od liczby wyróżnionych kategorii. Jeśli klasyfikacja zawodowa będzie bardziej szczegółowa to można oczekiwać, że odsetki te będą relatywnie mniejsze. Ponadto, odsetek osób dziedziczących pozycję zależy od różnic w strukturze zawodowej ojców i synów, czyli rozbieżności w rozkładach brzegowych obydwu zmiennych w tabeli 3.2. Zauważmy, że suma odsetków dziedziczenia może wynieść 100% tylko wówczas, jeśli odsetki te będą identyczne. Różnica w strukturze zawodowej ojców i synów powoduje, że pewna część osób „musi" być mobilna.

Do pomiaru wielkości różnic w strukturze można posłużyć się indeksem rozbieżności. W poprzednich rozdziałach służył on do określenia odstępstw danych z próby od modelu, w tym przypadku, będzie on odpowiadał na pytanie jaki odsetek mężczyzn powinien należeć do innej grupy zawodowej w tabeli 3.2 aby ich struktura zawodowa była identyczna ze strukturą zawodową ojców. Dla naszych danych wynosi on 24%,

tj:

$$\frac{|10,3-6,0|+|15,7-13,9|+|5,4-3,9|+|47,0-30,6|-|6,1-9,1|+|15,5-36,6|}{2}$$

gdzie 10,3% to odsetek synów należących do kategorii „inteligencja i wyższe kadry kierownicze" 6,0% to analogiczny odsetek wśród ojców. Innymi słowy aby respondenci niemobilni mogli stanowić 100% badanej próby 24% z nich należałoby zaklasyfikować do innej kategorii. Oznacza to, że zmiana struktury zawodowej jaka nastąpiła w pokoleniu synów w stosunku do pokolenia ojców[5], „wymusiła" pewną ruchliwość.

Z powyższych względów odsetki dziedziczenia niekoniecznie są dobrym miernikiem otwartości struktury społecznej. W literaturze rozróżnia się ruchliwość *absolutną* od ruchliwości *względnej*. Opisana powyżej miara odpowiada ruchliwości absolutnej, natomiast mierząc ruchliwość względną powinno się „kontrolować" różnice w rozkładach brzegowych. Mówiąc w pewnym uproszczeniu, badając ruchliwość względną chodzi o uzyskanie odpowiedzi na pytanie, jaka byłaby ruchliwość, gdyby nie występowały różnice w strukturze zawodowej ojców i synów. Większość badaczy jest zgodnych co do tego, że ruchliwość względna, a nie absolutna zdaje sprawę z otwartości struktury społecznej (dyskusja dotycząca tego zagadnienia zostanie przywołana w dalszej części tego rozdziału). Jak zostanie pokazane, właśnie modele logarytmiczno–liniowe pozwalają mierzyć natężenie ruchliwości przy kontroli różnic strukturalnych.

Kolejna tabela (3.5) również opisuje ruchliwość społeczną, przy czym tym razem pozycja ojca i syna porównywana jest pod kątem wykształcenia. W dalszej części będę nazywał ją w skrócie tablicą ruchliwości edukacyjnej. W odróżnieniu od wcześniejszej tabeli, kolejne kategorie tworzą pewną hierarchię, tj. można traktować je jak zmienne porządkowe[6]. Podobnie jak w odniesieniu do tablicy ruchliwości zawodowej również w odniesieniu do ruchliwości edukacyjnej można wyznaczyć odsetki napływu, odpływu i odsetki dziedziczenia. Przykładowo, odsetek osób niemobilnych: wynosi 34,1%.

[5] Określenie „pokolenie ojców" nie jest w pełni precyzyjne. Należy zauważyć, że rozkład zmiennej dla ojców, nie odnosi się do struktury społeczno-zawodowej mężczyzn w żadnym określonym momencie, w przeciwieństwie do zbiorowości respondentów. Różnice w rozkładach brzegowych mogą wynikać z wielu czynników demograficznych, nie tylko faktycznych zmian struktury społeczno-zawodowej (Duncan 1966, Kahl 1957, Boudon 1973, Matras 1961, Pullum 1975, Lissowski 1991).

[6] Trzeba zaznaczyć, że niektóre kategorie trudno ze sobą porównywać, na przykład kategorie *nieukończone średnie* i *zasadnicze zawodowe* albo *średnie zawodowe* i *średnie ogólnokształcące*. Zakładamy, że kategorie w tabeli 3.5 zostały połączone w taki sposób, że są one uporządkowane zgodnie z wymaganym poziomem wiedzy i umiejętności. Oczywiście, założenie to może wydawać się kontrowersyjne, trzeba również zwrócić uwagę na fakt, że uzyskane w ten sposób kategorie nie są homogeniczne. Pamiętając o powyższych zastrzeżeniach, dla celów prezentacji zmienne te będą traktowane jako przykłady zmiennych porządkowych.

Ze względu na to, że ojcowie i ich synowie porównywani są ze względu na zmienną porządkową, można dodatkowo określić, jaki odsetek synów posiada wykształcenie wyższe aniżeli ich ojcowie, a jaki odsetek posiada wykształcenie niższe. Odsetki te wynoszą odpowiednio: 59,1% (komórki powyżej przekątnej) oraz 6,7% (komórki poniżej przekątnej).

Tabela 3.5: Wykształcenie respondenta i wykształcenie jego ojca — tablica ruchliwości edukacyjnej[a]

| | Wykształcenie syna | | | | |
|---|---|---|---|---|---|
| Wykształcenie ojca | 1 | 2 | 3 | 4 | Suma |
| 1. Podstawowe i niepełne podstawowe | 343 | 502 | 199 | 101 | 1145 |
| 2. Niepełne średnie (w tym zasadnicze zawodowe) | 28 | 120 | 79 | 38 | 265 |
| 3. Ukończone średnie | 4 | 53 | 59 | 72 | 188 |
| 4. Niepełne wyższe i wyższe | 0 | 13 | 14 | 49 | 76 |
| Suma | 375 | 688 | 351 | 260 | 1674 |

[a]Źródło: Struktura społeczna II, 1987.

Tego rodzaju opis nie był możliwy w odniesieniu do tablicy ruchliwości zawodowej, gdyż przynależność społeczno–zawodowa jest cechą nominalną i kontrowersyjne jest porównywanie ze sobą różnych kategorii. Na przykład trudno jednoznacznie ocenić, czy syn rolnika, który został robotnikiem, ma wyższą, czy tez niższą pozycje zawodową względem swojego ojca[7]. Ponadto, w odniesieniu do wybranych osób, możliwe jest porównanie stopnia awansu bądź degradacji społecznej. Przykładowo: porównujemy dwie osoby, których ojcowie posiadali wykształcenie podstawowe. Jedna z nich ma wykształcenie niepełne średnie, druga ukończone wyższe. Choć obydwie te osoby „poprawiły" swoje wykształcenie w stosunku do wykształcenia ich ojca, to przypadku drugiego z nich „awans" ten był większy. Korzystając z porządkowego charakteru analizowanych zmiennych można określić o ile kategorii (jedną, dwie, bądź więcej) w „górę" bądź w „dół" różni się wykształcenie badanej osoby w stosunku do pozycji ich

[7]W literaturze zawód jest często traktowany jako zmienna porządkowa, tj. zakłada się występowanie pewnej hierarchii, pomiędzy poszczególnymi grupami zawodowymi (Hout 1983). Czasem badanie ruchliwości ogranicza się jedynie do tych grup zawodowych, dla których ustalenie takiego uporządkowania ma mocniejsze podstawy teoretyczne (Glass 1954, Miller 1960, Duncan 1979)

ojca. Odwołując się do zdefiniowanego wcześniej pojęcia: im dalej pseudo–przekątna oddalona jest od głównej przekątnej, tym awans (bądź degradacja) dotyczy większej liczby kategorii. Jak zostało zasygnalizowane wcześniej rozróżnienia tego typu — wyróżnienie kolejnych pseudo–przekątnych — pozwoli na formułowanie dodatkowych modeli, których nie można formułować w odniesieniu do zmiennej nominalnej.

Podobnie jak w przypadku ruchliwości zawodowej na odsetki opisujące ruchliwość edukacyjną wpływ mają różnice w rozkładach brzegowych tj. rozbieżności w strukturze wykształcenia ojców i synów. Dotyczy to odsetków dziedziczenia, ale również innych aspektów ruchliwości. Indeks rozbieżności wynosi prawie 46%. Ponieważ obydwie zmienne mierzone są na skali porządkowej możliwe jest zastosowanie innego wskaźnika tzw. skali rozbieżności (Lieberson 1976, Fossett i inni 1986), który będzie w tej pracy oznaczony jako ND (*index of net difference*). Miernik ten dany jest wzorem:

$$ND = \sum_{i=2}^{r} P(X = x_i)P(Y < y_i) - \sum_{i=2}^{r} P(Y = y_i)P(X < x_i) \tag{3.1}$$

Miernik ten porównuje rozkłady brzegowe obydwu zmiennych. Jego interpretacja jest następująca: przypuśćmy, że wyznaczamy rozkład łączny zmiennych w oparciu o te rozkłady i założenie niezależności, tj. prawdopodobieństwo każdej kombinacji obydwu zmiennych jest iloczynem odpowiednich prawdopodobieństw brzegowych. Wówczas wyrażenie $\sum_{i=2}^{r} P(X = x_i)P(Y < y_i)$ wskazuje na prawdopodobieństwo, że zmienna $X$ jest większa od zmiennej $Y$, a wyrażenie $\sum_{i=2}^{r} P(Y = y_i)P(X < x_i)$, że zmienna $Y$ jest większa od zmiennej $X$. Gdyby rozkłady były identyczne obydwa wyrażenia byłyby sobie równe i w konsekwencji miernik $ND$ wynosiłby 0. Jeśli miernik jest większy od 0 wskazuje to, że „wysokie" wartości zmiennej $X$ występują relatywnie częściej niż „wysokie" wartkości zmiennej $Y$, wartość ujemna na sytuację odwrotną.

W przypadku tablicy ruchliwości edukacyjnej $ND = -46,1\%$, co wskazuje, że wśród synów kategorie opisujące wyższe kategorie wykształcenia są częściej reprezentowane niż wśród ich ojców. Dokładniej, taka byłaby różnica pomiędzy odsetkiem respondentów, którzy mają wykształcenie niższe od swojego ojca a odsetkiem respondentów, którzy mają od swojego ojca wykształcenie wyższe, jeśli wziąć pod uwagę rozkłady brzegowe obydwu zmiennych i przyjąć założenie niezależności obydwu zmiennych. Wielkość tę można zestawić z tym co faktycznie dzieje się w tablicy ruchliwości edukacyjnej: tam różnica pomiędzy odsetkiem osób poniżej i powyżej przekątnej wynosi 52,5%, tak jej wartość absolutna jest większa. Generalnie częściej zdarza się awans społeczny respondenta w stosunku do pozycji ojca, aniżeli degradacja. W pewnej mierze daje się to wyjaśnić zmianami strukturalnymi (na co wskazuje indeks ND), ale nie do końca, co pokazuje porównanie obydwu powyższych wielkości. W dalszej części tego rozdziału, zaprezentowane modele logarytmiczno–linowe

dotyczące asymetrii. Pozwolą one odpowiedzieć na pytanie czy asymetrię daje się zaobserwować również przy kontroli różnic w rozkładach brzegowych.

Zanim przedstawione zostaną kolejne przykłady danych o takich samych kategoriach analizowanych zmiennych, warto poświęcić kilka słów badaniom dotyczącym ruchliwości społecznej. Ruchliwość stanowi niezwykle ważne zagadnienie w socjologii, które traktuje się je jako przejaw otwartości społecznej, wskazuje na „równość szans" jednostek. Ponadto jest istotnym czynnikiem wpływającym na kształt struktury społecznej, np. formowania się struktury klasowej. Można rozpatrywać ruchliwość pod kątem wpływu na postawy społeczne i psychologiczne charakterystyki jednostek, z drugiej strony można analizować wpływ ruchliwości na funkcjonowanie gospodarki, czy też występowanie konfliktów społecznych (Domański 2007*b*).

Za twórcę definicji ruchliwości uważa Pitrima Sorokina (1927), podjął on jednocześnie próbę wielowymiarowego wyjaśnienia mechanizmów z nią związanych, ponadto przyczynił się sformułowania wielu hipotez związanych z tym zagadnieniem. Nie sposób w pracy tego typu prześledzić szczegółowo badań dotyczących analizy tego zjawiska i różnych jego aspektów. Warto zaznaczyć, że wyróżniono wiele obszarów badawczych, między innymi rozróżnia się ruchliwość horyzontalną od ruchliwości pionowej. Pierwsza z nich — w odróżnieniu od drugiej — opisuje zmianę pozycji nie związaną bezpośrednio ze zmianą położenia jednostki w hierarchii społecznej, np. zmianę wyznania, miejsca zamieszkania, itp. Jak widać z zamieszonych powyżej przykładów tablic ruchliwości zawodowej i edukacyjnej, w tej pracy analizowana jest raczej ruchliwość pionowa[8].

Dla celów tej pracy istotne jest również przytoczone powyżej wyróżnienie różnych aspektów ruchliwości absolutnej, tj. ruchliwości strukturalnej — związanej ze zmianą struktury zawodowej lub struktury wykształcenia, przykładowo pomiędzy pokoleniem rodziców i ich dzieci — i ruchliwości względnej, która często znajduje się w centrum zainteresowania badaczy struktury społecznej. Interesującą pracą podkreślającą wagę tego rozróżnienia był artykuł Kazimierza Słomczyńskiego i Tadeusza Krazue (1986). Bardziej szczegółową dekompozycję tablicy ruchliwości prezentuje Grzegorz Lissowski (1991).

Zjawisko ruchliwości analizowane jest za pomocą wielu metod statystycznych. Wykorzystuje się do tych celów między innymi: skalowanie wielowymiarowe (Blau i Duncan 1978), analizę kanoniczną (Klatzky i Hodge 1971, Domański i Sawiński 1987), analizę dyskryminacyjną i analizę korespondencji. Jak zostanie pokazane w dalszej

[8]W socjologii zmiany tego typu zwykło się interpretować jako społeczny awans lub degradację. Jednak, jak zostało zasygnalizowane wcześniej, nie wszystkie zmiany tego typu da się w ten sposób jednoznacznie określić, co wynika z tego, że przynależność społeczno–zawodowa nie jest zmienną porządkową.

części tej pracy potrzeba rozróżnienia ruchliwości strukturalnej od ruchliwości względnej przesądziła o szerokim wykorzystaniu modelowania logarytmiczno–liniowego. W modelach tych bowiem rozróżnia się efekty związane z efektami poszczególnych zmiennych od parametrów opisujących wzór i siłę związku między zmiennymi, dzięki temu wykorzystuje się te ostatnie do opisu ruchliwości względnej.

Przykładowo, istotnym zagadnieniem interesującym socjologów była kwestia pomiaru siły dziedziczenia poszczególnych kategorii zawodowych lub kategorii wykształcenia. Wielu badaczy próbowało odpowiedzieć na pytanie jaka jest tendencja do dziedziczenia tej samej pozycji, jeśli kontroluje się różnice zmiany strukturalne. Proponowano różne mierniki bądź dyskutowano ich własności (m. in. Rogoff 1953, Glass 1954, Tumin i Feldman 1957, Yasuda 1964, Duncan 1966, Wilensky 1966, Duncan 1966, Goodman 1961, 1965, 1969, Featherman i Hauser 1978, Hope 1981, Sawiński 1981). Przykładowo, wartość powszechnie wykorzystywanego indeksu zaproponowanego przez Glassa i Rogoff jest w pewnej mierze zależna od rozkładów brzegowych. Obecnie powszechnie wykorzystuje się mierniki zaproponowane przez Goodmana, które opierają się właśnie na modelowaniu logarytmiczno–liniowym.

Warto podkreślić, że rozwój modeli logarytmiczno–liniowych, w dużej mierze był inspirowany właśnie przez badania nad ruchliwością społeczną (Goodman 1972a, 1979a, 1979b, Hauser 1980, Hout 1983, Yamaguchi 1987, Sobel i inni 1985, Xie 1992, Goodman i Hout 1998). Z drugiej strony wiele badań dotyczących ruchliwości wykorzystuje właśnie tę metodę analityczną, (m.in Featherman i inni 1975, Erikson i Goldthorpe 1992, Breen 2006).

Przy tej okazji warto wskazać również prace polskich autorów. Kwestię ruchliwości analizowano zanim w Polsce zaczęto posługiwać się technika modelowania logarytmiczno–liniowego (m. in. Sarapata 1965, Janicka 1973, 1976, Słomczyński 1973). Warto też przywołać badanie ogólnopolskie, kierowane przez Michała Pohoskiego w 1972 roku, które pozwoliło na wnikliwe przeanalizowanie tej i innych kwestii dotyczących struktury społecznej[9] . Trudno wymienić wszystkie prace polskich socjologów dotyczących ruchliwości, warto jednak zauważyć, ze wiele z nich wykorzystywało modelowanie logarytmiczno–liniowe (m. in. Pohoski 1983, Mach 2002, 2004, Domański 2004, Domański i inni 2008). Warto w tym miejscu przywołać prace o nachyleniu metodologicznym Andrzeja Kutyłowskiego (m. in. 1988).

Tabela 3.6 jest przykładem danych panelowych. Dane dotyczą gospodarstw domowych i pochodzą z I i III edycji badania „Diagnoza społeczna", które odbyły się odpowiednio w 2000 i 2005 roku[10] (Czapliński i Panek 2007). Badani mieli określić sy-

[9]Badanie to było kontynuowane, kolejne edycje odbyły się w 1987, 1991 i 1998 roku.

[10]Dane zostały przeważone, dlatego liczebności w tabeli nie są liczbami całkowitymi.

tuację finansową swoich gospodarstw domowych odpowiadając na pytanie „Czy przy aktualnym dochodzie netto Pana(i) gospodarstwo domowe wiąże koniec z końcem?" Ankietowani mogli wybrać jedną z odpowiedzi wymienionych w tabeli 3.6. Porównane zostały odpowiedzi udzielone w obydwu badaniach.

Tabela 3.6: Ocena sytuacji materialnej gospodarstwa domowego w 2000 i 2005 roku[a]

| Czy przy aktualnym dochodzie netto Pana(i) gospodarstwo domowe wiąże koniec z końcem? | | | | | | |
|---|---|---|---|---|---|---|
| Odpowiedzi w 2000 roku ($X$) | Odpowiedzi w 2005 roku ($Y$) | | | | | |
| | 1 | 2 | 3 | 4 | 5 | Suma |
| 1. Z wielką trudnością | 271,7 | 171,0 | 112,5 | 24,1 | 2,9 | 582,1 |
| 2. Z trudnością | 100,2 | 134,5 | 169,8 | 45,2 | 6,8 | 456,5 |
| 3. Z pewną trudnością | 58,0 | 131,3 | 248,5 | 97,3 | 10,8 | 545,9 |
| 4. Raczej łatwo | 8,6 | 27,1 | 81,8 | 73,9 | 22,6 | 214,0 |
| 5. Łatwo | 1,2 | 2,1 | 7,8 | 12,6 | 13,5 | 37,3 |
| Suma | 439,7 | 465,9 | 620,4 | 253,1 | 56,7 | 1835,8 |

[a]Źródło: Diagnoza społeczna, 2000-2005, dane przeważone.

Struktura tej tabeli jest pod wieloma względami podobna jak tabeli ruchliwości. Przekątna tej tabeli opisuje gospodarstwa, których członkowie nie zmienili oceny sytuacji materialnej, a dokładniej ocena ta jest taka sama w obydwu badaniach[11]. Odsetek ten wynosi 40,4%. Powyżej przekątnej są gospodarstwa (36,1%), których ocena dotycząca sytuacji materialnej w 2005 roku jest lepsza aniżeli w roku 2000. Komórki poniżej przekątnej (23,5%) analogicznie opisują gospodarstwa, w których sytuacja materialna w roku 2005 pogorszyła się, jeśli porównać ją do 2000 roku. Tak jak poprzednio, na poszczególne odsetki mają wpływ różnice w rozkładach brzegowych. Nie są one jednak tak duże jak w przypadku tablic ruchliwości, niemniej rozkład zmiennej opisującej ocenę sytuacji finansowej zmienił się w ciągu pięciu lat: indeks rozbieżności wynosi $\Delta = 7,8\%$, a skala rozbieżności $ND = -9,8\%$.

Poniżej przedstawione zostaną trzy inne przykłady danych o takiej samej strukturze jak tablice ruchliwości i dane panelowe, mianowicie kategorie analizowanej zmiennej wierszowej i kolumnowej są identyczne. Pierwszy z nich przedstawia tabela 3.7 i

[11]Nie można wykluczyć, że sytuacja osoby deklarującej w 2000 i 2005 roku, że jej gospodarstwo wiązało koniec z końcem „z wielką trudnością" mogła zmieniać się w analizowanym okresie, przykładowo w roku 2002 mogłaby odpowiedzieć „łatwo". W tym sensie sformułowanie o „braku zmian w czasie" nie jest w pełni precyzyjne, gdyż obserwujemy odpowiedzi respondenta jedynie w dwóch punktach czasowych.

dotyczy wzorów zawierania małżeństw. Tabela opisuje małżeństwa zawarte w 2003 roku i opiera się na danych urzędowych zbieranych przez Główny Urząd Statystyczny (2004). Wzory zawierania małżeństw badane z perspektywy wykształcenia bądź zawodu, podobnie jak ruchliwość społeczną uważa się za jeden z głównych wskaźników otwartości struktury społecznej. Na ogół zjawisko to bada się za pomocą tych samych modeli logarytmiczno–liniowych, jakie stosuje się do badania ruchliwości (m. in. Mare 1991, Pohoski 1991, Kalmijn 1991, Smits i inni 1998, Halpijn i Chan 2003, Domański i Przybysz 2007).

Tabela 3.7: Małżeństwa zawarte w 2003 roku według wykształcenia męża i żony[a]

| Wykształcenie męża | Wykształcenie żony | | | | | | Suma |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| 1. Niepełne podstawowe | **68** | 69 | 28 | 25 | 1 | 1 | 192 |
| 2. Podstawowe | 63 | **5861** | 5409 | 5068 | 158 | 540 | 17099 |
| 3. Zasadnicze zawodowe | 38 | 7231 | **22606** | 27402 | 941 | 4854 | 63072 |
| 4. Średnie | 5 | 2949 | 8441 | **42771** | 1344 | 17056 | 72566 |
| 5. Policealne | 0 | 26 | 120 | 506 | **301** | 402 | 1355 |
| 6. Wyższe | 0 | 207 | 720 | 10270 | 657 | **25640** | 37494 |
| Suma | 174 | 16343 | 37324 | 86042 | 3402 | 48493 | 191778 |

[a]Źródło: Rocznik Demograficzny, GUS (2004), s. 256.

W kolejnych wierszach podane zostały kategorie wykształcenia męża, w kolejnych kolumnach przedstawione są kategorie wykształcenia żony. Na przekątnej tabeli są małżeństwa, w których wykształcenie męża jest takie samo jak wykształcenie żony, czyli tzw. małżeństwa homogeniczne. Stanowią one ponad połowę (50,7%) wszystkich małżeństw zawartych w 2003 roku. Jeśli chodzi o małżeństwa heterogeniczne to można je podzielić na te w których lepiej wykształcona jest żona (komórki powyżej przekątnej) oraz na małżeństwa, w których lepiej wykształcony jest mąż (komórki poniżej głównej przekątnej). Ich odsetki wynoszą odpowiednio 33% oraz 16,3%. Podobnie jak w przypadku ruchliwości niedopasowanie struktury wykształcenia mężów i żon „wymusza" pojawianie się małżeństw heterogenicznych. W przypadku analizowanych danych indeks rozbieżności wynosi 13,8% a skala rozbieżności jest równa $ND = -28,6\%$. Ujemny znak drugiego miernika wskazuje, że wśród kobiet wyższe kategorie wykształcenia zdarzają się relatywnie częściej niż wśród mężczyzn. Kon-

trola tych różnic — na co pozwalają modele logarytmiczno–liniowe — pozwala na zbadanie faktycznej tendencji do homogamii bądź asymetrii.

Choć tablice ruchliwości są pod wieloma względami podobne do tabel opisujących wzory zawierania małżeństw, warto zwrócić uwagę na kilka różnic. Część osób nie zawiera związku małżeńskiego. Chcąc interpretować wzory zawierania małżeństw jako wskaźnik otwartości struktury społecznej, należałoby również uwzględnić ten fakt. Na przykład, jeśli kobiety z wyższym wykształceniem relatywnie częściej pozostają samotne w porównaniu do kobiet reprezentujących inne kategorie wykształcenia, może to wskazywać na istnienie silnych barier społecznych, tj. można podejrzewać, że niemożność znalezienia męża o podobnym statusie zmniejsza prawdopodobieństwo zamążpójścia. Innymi słowy, ze względu na to, że cześć osób w populacji nie zawiera związku małżeńskiego, interpretacja różnic w rozkładach brzegowych tabeli jako różnic w strukturze wykształcenia jest bardziej problematyczna niż w przypadku tablic ruchliwości.

Tabela 3.8: Opinie dotyczące przyjazdu do polski imigrantów z biedniejszych i botszych krajów europejskich[a]

| Powinno się zezwalać na przyjazd ... | | | | | |
|---|---|---|---|---|---|
| ...z bogatszych krajów europejskich | ...z biedniejszych krajów europejskich | | | | |
| | 1 | 2 | 3 | 4 | Suma |
| 1. Zezwalać dużej liczbie osób | 154,6 | 125,7 | 55,9 | 5 | 341,2 |
| 2. Zezwalać pewnej liczbie osób | 29,8 | 741,9 | 219,7 | 5,0 | 1009,2 |
| 3. Zezwalać tylko nielicznym | 10,0 | 84,7 | 363,4 | 17,9 | 485,0 |
| 4. Nie zezwalać nikomu | 4,9 | 14,8 | 26,6 | 27,0 | 113,8 |
| Suma | 199,3 | 967,0 | 665,6 | 54,9 | 1949,2 |

[a]Źródło: Europejski Sondaż Społeczny, 2002, dane przeważone.

Na koniec podane zostaną przykłady dotyczące dwóch pytań zadanych respondentom, na które mogli odpowiedzieć posługując się tymi samymi kategoriami. W pierwszym przykładzie (tabela 3.8) respondentom zadano pytania dotyczące tego na ile państwo polskie powinno pozwalać na przyjazd do Polski osobom z biedniejszych

i bogatszych krajów europejskich (ESS, 2002). Kategorie odpowiedzi podane zostały w tabeli 3.8[12] Daje się zauważyć wiele podobieństw w analizie tablicy tego typu do analizy tablic ruchliwości i danych panelowych. Odpowiedzi na głównej przekątnej wskazują na tych respondentów, którzy uważają, że imigrantom z biedniejszych i bogatszych krajów europejskich powinno się zezwalać na przyjazd do Polski w takim samym zakresie.

Tabela 3.9: Czas poświęcony na oglądanie telewizji i słuchanie radia w dzień powszedni[a]

| Ile czasu poświęca Pan(i) w typowy dzień powszedni na ... | | | | | | |
|---|---|---|---|---|---|---|
| oglądanie telewizji | słuchanie radia | | | | | Suma |
| | 1 | 2 | 3 | 4 | 5 | |
| 1. W ogóle | 18,9 | 15,9 | 8,0 | 8,8 | 25,9 | 77,5 |
| 2. Do 1 godziny | 88,2 | 139,0 | 68,1 | 40,2 | 131,5 | 466,8 |
| 3. Do 2 godzin | 139,1 | 209,4 | 81,5 | 76,6 | 160,1 | 666,8 |
| 4. Do 3 godziny | 104,6 | 148,9 | 68,7 | 40,8 | 127,1 | 490,1 |
| 5. Ponad 3 godziny | 112,3 | 97,5 | 55,3 | 28,9 | 98,8 | 392,9 |
| Suma | 439,7 | 465,9 | 620,4 | 253,1 | 56,7 | 2094,1 |

[a] Źródło: Europejski Sondaż Społeczny, 2002, dane przeważone.

Dane o podobnej strukturze pochodzące z tego samego badania przedstawione są w tabeli 3.9. Zamieszczono w niej odpowiedzi na pytanie o czas poświęcany na oglądanie telewizji i słuchanie radia[13]. Jest to przykład danych, który pokazuje, że nie zawsze komórki zamieszczone na głównej przekątnej — co w tym przypadku oznacza taką samą ilość czasu przeznaczoną na obydwie te czynności — charakteryzują się wysokimi odsetkami. Dla analizowanej tabeli obejmują one jedynie 18% próby. Wynika to w pewnej mierze z tego, że trudno wskazać silne argumenty przemawiające za tym, że częste oglądanie telewizji idzie w parze z częstym słuchaniem radia, co więcej ograniczenia czasowe sprawiają, że stosunkowo trudno poświęcić na obydwie te czynności dużo czasu.

[12] W badaniu tym 7,6% udzieliło odpowiedzi „trudno powiedzieć" na co najmniej jedno z pytań. Osoby te zostały wyłączone z analizy.

[13] W badaniu tym mniej niż 1 procent udzieliło odpowiedzi „trudno powiedzieć" na co najmniej jedno z pytań. Osoby te zostały wyłączone z analizy.

## 3.2 Modele dla dwóch zmiennych o takich samych kategoriach

Jak zostało zasygnalizowane w poprzedniej części, analiza odsetków w odniesieniu do tablic ruchliwości i danych panelowych posiada istotne ograniczenia. Tak jak zostało zasygnalizowane wcześniej, nie można traktować odsetka respondentów mobilnych jako miary ruchliwości względnej, gdyż na wielkość tę istotny wpływ mają zmiany strukturalne, tj. różnice w rozkładach wykształcenia synów i ojców. Mówiąc inaczej aby opisać związek i siłę związku między zmiennymi należy „kontrolować" różnice w rozkładach brzegowych analizowanych zmiennych. Umożliwia to modelowanie logarytmiczno–liniowe, gdyż — jak zostało przedstawione w poprzednich rozdziałach — rozkład łączny dwóch zmiennych opisuje się z jednej strony za pomocą efektów poszczególnych zmiennych, z drugiej strony za pomocą parametrów interakcji. Dodatkowo, w ramach tej metody możliwe jest formułowanie hipotez dotyczących ruchliwości, związku pomiędzy zmiennymi przy analizie danych panelowych, itd. jak również ich weryfikacja. Modele logarytmiczno–liniowe nie są więc jedynie metodą opisową, jak było w przypadku charakterystyki danych za pomocą omawianych wcześniej odsetków (dziedziczenia, napływu, odpływu, asymetrii).

W tej części przedstawione zostaną najczęściej formułowane hipotezy dotyczące tablic ruchliwości i danych panelowych. Rozpoczniemy od modeli niezależności, quasi–niezależności, symetrii i quasi–symetrii. Te modele mogą być stosowane zarówno w odniesieniu do zmiennych nominalnych jak i porządkowych. Większość modeli prezentowanych w dalszej części tego rozdziału będzie wykorzystywało informację o uporządkowaniu kategorii analizowanych zmiennych. Przytoczone zostaną zaprezentowane już w rozdziale poprzednim modele jednakowej interakcji i modele wierszowo–kolumnowe. Tym razem omówione zostaną możliwości ich zastosowania do analizy tablic ruchliwości i danych panelowych, jak również przedstawione zostaną specyficzne modyfikacje tych modeli w kontekście omawianych danych. W dalszej części przedstawione zostaną modele dystansu i przekraczania barier, które — w przeciwieństwie do modeli jednakowej interakcji i modeli wierszowo–kolumnowych — wykorzystywane są praktycznie wyłącznie w odniesieniu do analizowanych zmiennych o takich samych kategoriach .

### 3.2.1 Niezależność i quasi–niezależność stochastyczna

Jako punkt wyjścia w analizach nad ruchliwością i danymi panelowymi przyjmuje się często model niezależności stochastycznej — będziemy go oznaczać jako N. Zgodnie z tą hipotezą wszystkie stosunki szans są równe 1. Wynikająca z tego modelu identycz-

ność rozkładów warunkowych jednej zmiennej względem drugiej zmiennej wskazuje — w odniesieniu do tablic ruchliwości — na brak związku między pozycją społeczną syna i pozycją społeczną ojca, tj. model taki zakłada, że w społeczeństwie nie istnieją bariery bądź przywileje, wynikające z pochodzenia społecznego, które utrudniałyby bądź ułatwiały uzyskanie odpowiedniej pozycji społecznej (wykształcenia bądź pozycji zawodowej). W przypadku danych panelowych oznaczałoby to, że cecha zaobserwowana w punkcie czasowych $t_2$ nie jest w żadnym stopniu zależna od tej cechy respondenta w obserwowanej w punkcie $t_1$.

Fakt, że obydwie zmienne mają identyczne kategorie pozwala na uwzględnienie dodatkowych założeń. Przykładowo, można uprościć model niezależności, dodając założenie o identyczności rozkładów zmiennej wierszowej i kolumnowej, tj:

$$\pi_i^X = \pi_i^Y,$$

dla każdego $i$, takiego, że $1 \leq i \leq r$. Model taki będzie oznaczany jako NS. Należy podkreślić, że związane z nim założenia są bardzo restrykcyjne i na ogól model taki nie jest realistycznym opisem danych. Pokazuje on jednak, w jaki sposób uwzględnienie identyczności kategorii obydwu zmiennych pozwala na wprowadzanie dodatkowych założeń, ponadto będzie on modelem odniesienia dla pozostałych hipotez przedstawionych w tym rozdziale. Każda z kolejnych hipotez będzie bardziej złożona od modelu NS, mówiąc inaczej model ten będzie zagnieżdżony[14] w każdym z kolejnych prezentowanych modeli. Tabela 3.10 przedstawia ilustrację tego rozkładu. W tabeli tej wielkości $d_i$ można interpretować jako parametry opisujące efekt poszczególnych zmiennych. Parametry interakcji są równe 1, tak jak w modelu niezależności, więc nie zostały uwzględnione w tabeli. Parametryzacja tego modelu wygląda następująco:

$$\pi_{ij}^{XY} = d \cdot d_i \cdot d_j \tag{3.2}$$

Parametry $d_i$, $d_j$ nie posiadają indeksów górnych wskazujących, której zmiennej dotyczą, gdyż zachodzi $d_i^X = d_i^Y$. Ujmując to inaczej, w stosunku do modelu niezależności model NS zakłada dodatkowo równość parametrów opisujących efekty poszczególnych zmiennych, o ile dotyczą one tej samej kategorii zmiennej wierszowej i kolumnowej. Model ten posiada więc 1 parametr $d$ i $(r-1)$ niezależnych parametrów $d_i$. Jego liczba stopni swobody wynosi:

$$df = (r^2 - 1) - (r - 1) = r(r - 1).$$

[14]Definicja modelu zagnieżdżonego została przedstawiona w rozdziale pierwszym, przypomnijmy, jedynie, że model A jest zagnieżdżony w modelu B jeśli jest od niego prostszy, tj. można go uzyskać dodając do modelu B dodatkowe założenia.

W stosunku do modelu niezależności liczba ta jest większa o $(r-1)$, co wynika z założenia, że rozkład zmiennej wierszowej o $r$ kategoriach jest taki sam, jak rozkład zmiennej $Y$. Jak zostało przedstawione w rozdziale pierwszym, wyznaczając metodą największej wiarygodności rozkład oczekiwany zgodny z modelem niezależności należało przyjąć, że rozkłady brzegowe są identyczne z rozkładem z próby. W przypadku modelu NS zakłada się dodatkowo, że częstości brzegowe obydwu zmiennych są identyczne, tak więc przy estymacji musi być spełniony dodatkowy warunek:

$$\widehat{\pi}_i^X = \widehat{\pi}_i^Y = \frac{p_i^X + p_i^Y}{2}.$$

Tak więc zgodnie z metodą największej wiarygodności brzegowe prawdopodobieństwo $i-$tej kategorii zmiennej $X$ i zmiennej $Y$ szacowane jest jako średnia dwóch brzegowych empirycznych częstości obydwu zmiennych dotyczących tej kategorii.

Tabela 3.10: Ilustracja modelu niezależności NS — parametry opisujące efekty główne zmiennych $X$ i $Y$

| $X\backslash Y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $y_1$ | $d_1^2$ | $d_1d_2$ | $d_1d_3$ | $d_1d_4$ | $d_1d_5$ |
| $y_2$ | $d_2d_1$ | $d_2^2$ | $d_2d_3$ | $d_2d_4$ | $d_2d_5$ |
| $y_3$ | $d_3d_1$ | $d_3d_2$ | $d_3^2$ | $d_3d_4$ | $d_3d_5$ |
| $y_4$ | $d_4d_1$ | $d_4d_2$ | $d_4d_3$ | $d_4^2$ | $d_4d_5$ |
| $y_5$ | $d_5d_1$ | $d_5d_2$ | $d_5d_3$ | $d_5d_4$ | $d_5^2$ |

Jak się łatwo domyślić obydwa powyższe modele N oraz NS na ogół nie są adekwatnym opisem tablic ruchliwości i danych panelowych. Statystyki $X^2$ oraz $G^2$ zostały przedstawione w tabeli 3.11. Na przykład, dla tablicy ruchliwości zawodowej (3.2) statystyka $G^2$ wynosi 432,0 ($p < 0,0001$) dla modelu niezależności, a w przypadku modelu, który dodatkowo zakłada identyczność struktury zawodowej ojca i syna jest ona równa 661,3 ($p < 0,0001$). Indeks rozbieżności dla modelu niezależności wynosi 19,4. Wielkość ta wskazuje, jaki odsetek osób należałoby przesunąć w tabeli rozkładu łącznego, aby dane te były zgodne z modelem. Dla modelu NS odsetek ten wynosi 24,8%. Do podobnych konkluzji prowadzą wyniki dopasowania dla tablicy ruchliwości edukacyjnej i danych panelowych porównujących ocenę sytuacji materialnej własnego gospodarstwa domowego w 2000 i 2005 roku.

Modele te, choć nierealistyczne, traktuje się na ogół jako modele odniesienia. Przypomnijmy, że w kontekście ruchliwości model niezależności wskazuje na hipotetyczną sytuację braku barier społecznych, które przekładałyby się na łatwość uzyskania pozycji zawodowej lub wykształcenia. Z kolei model NS jest najprostszym z modeli

Tabela 3.11: Wyniki weryfikacji hipotez dla tabel 3.2, 3.5, 3.6

| Model | df | $\chi^2$ | $G^2$ | $\Delta$ |
|---|---|---|---|---|
| Wyniki dla tabeli 3.2 (tablica ruchliwości zawodowej) | | | | |
| NS | 30 | 616,3 ($p < 0,0001$) | 661,3 ($p < 0,0001$) | 24,8 |
| N | 25 | 482,9 ($p < 0,0001$) | 432,0 ($p < 0,0001$) | 19,4 |
| QN | 19 | 115,3 ($p < 0,0001$) | 104,2 ($p < 0,0001$) | 7,8 |
| QhN | 24 | 186,7 ($p < 0,0001$) | 184,1 ($p < 0,0001$) | 13,2 |
| QNS | 24 | 439,3 ($p < 0,0001$) | 441,6 ($p < 0,0001$) | 16,0 |
| S | 15 | 318,1 ($p < 0,0001$) | 362,6 ($p < 0,0001$) | 14,7 |
| QS | 10 | 20,7 ($p = 0,0235$) | 18,0 ($p = 0,0556$) | 2,4 |
| Wyniki dla tabeli 3.5 (tablica ruchliwości edukacyjnej) | | | | |
| NS | 12 | 972,4 ($p < 0,0001$) | 1097,6 ($p < 0,0001$) | 28,8 |
| N | 9 | 359,0 ($p < 0,0001$) | 337,8 ($p < 0,0001$) | 15,1 |
| QN | 5 | 123,4 ($p < 0,0001$) | 105,1 ($p < 0,0001$) | 6,3 |
| QhN | 8 | 236,5 ($p < 0,0001$) | 226,8 ($p < 0,0001$) | 12,8 |
| QNS | 8 | 900,6 ($p < 0,0001$) | 1028,7 ($p < 0,0001$) | 26,4 |
| S | 6 | 768,7 ($p < 0,0001$) | 958,4 ($p < 0,0001$) | 26,3 |
| QS | 3 | 7,4 ($p = 0,0593$) | 8,1 ($p = 0,0439$) | 1,1 |
| Wyniki dla tabeli 3.6 (dane panelowe) | | | | |
| NS | 20 | 632,2 ($p < 0,0001$) | 559,2 ($p < 0,0001$) | 21,1 |
| N | 16 | 617,0 ($p < 0,0001$) | 527,1 ($p < 0,0001$) | 19,9 |
| QN | 11 | 221,2 ($p < 0,0001$) | 185,4 ($p < 0,0001$) | 9,1 |
| QhN | 15 | 326,3 ($p < 0,0001$) | 323,5 ($p < 0,0001$) | 16,9 |
| QNS | 15 | 274,3 ($p < 0,0001$) | 235,4 ($p < 0,0001$) | 10,0 |
| S | 10 | 60,5 ($p < 0,0001$) | 61,6 ($p < 0,0001$) | 6,3 |
| QS | 6 | 1,5 ($p = 0,9624$) | 1,4 ($p = 0,9636$) | 0,7 |

opisywanych w tym rozdziale. W kolejnych z prezentowanych modeli uchylone zostanie założenie identyczności rozkładów brzegowych bądź braku związku między zmiennymi.

Często proponowaną przez badaczy modyfikacją jest uwzględnienie tego, że zarówno w odniesieniu do tablic ruchliwości, jak też danych panelowych, komórki na przekątnej opisujące współwystępowanie tej kategorii obydwu zmiennych występują relatywnie częściej niż wynika to z modelu niezależności. Przypomnijmy, że można to

interpretować jako tendencję do dziedziczenia pozycji społecznej bądź występowania tej samej kategorii zmiennej w punktach czasowych $t_1$ i $t_2$.

W 1955 roku Blumen i inni (1955) zaproponowali model podziału całej populacji na osoby „mobilne" i „niemobilne" (*mover-stayer model*). Zgodnie z tą koncepcją w populacji można wyróżnić:

1. osoby, dla których obydwie zmienne mają taką samą wartość obydwu obserwowanych zmiennych (*stayers*). W odniesieniu do problemu ruchliwości międzypokoleniowej byłyby to osoby, które miałyby taką samą pozycję społeczną jak ich ojciec,

2. osoby, dla których wartości obydwu obserwowanych zmiennych są różne (*movers*). W odniesieniu do problemu ruchliwości byłyby to osoby, które mają inną pozycję społeczną aniżeli ich ojciec.

Goodman (1961) przeformułował powyższą koncepcję. Zgodnie z zaproponowanym przez niego modelem wzór zależności w tablicy ruchliwości jest wynikiem dwóch procesów.

1. „nie–mobilności", tj. pewien odsetek osób ma taką samą wartość obydwu analizowanych zmiennych (zajmuje tę samą pozycję społeczną)

2. „mobilności", tj. dla pozostałych osób wartości obydwu zmiennych nie są od siebie zależne, np. pozycja społeczna zajmowana przez syna nie zależy od pozycji ojca. Podkreślić należy, że niezależność nie oznacza zależności negatywnej, tj. *nie wyklucza się*, że będzie to taka sama pozycja.

Warto już w tym miejscu zauważyć, że model ten bywa uogólniany (Becker 1990), tj. mobilność nie musi być opisywana za pomocą niezależności dwóch zmiennych, ale może być to inny typ zależności. Taki ogólny model będzie oznaczany jako $MS$. Przykłady bardziej ogólnego rozumienia mobilności podane zostaną w dalszej części tej pracy, prezentację warto jednak rozpocząć od oryginalnej wersji, zgodnie z którą, dla osób mobilnych zmienna wierszowa i kolumnowa są niezależne — oznaczmy ten model jako ($MS_0$). Zgodnie z tym modelem, o ile osoby niemobilne znajdują się wyłącznie na głównej przekątnej, to pozostałe osoby lokują się na głównej przekątnej jak również poza nią. Można to przedstawić następująco[15]: zmienna $Z$ określa, czy dana osoba należy do populacji osób niemobilnych ($Z = 1$), czy też mobilnych ($Z = 2$). Odsetki tych podzbiorowości wynoszą odpowiednio $\beta_1$ oraz $\beta_2$ i sumują się do 1. W takim ujęciu prawdopodobieństwa rozkładu łącznego wynoszą:

$$\pi_{ij}^{XY} = \pi_{ij(1)}^{XY(Z)}\beta_1 + \pi_{ij(2)}^{XY(Z)}\beta_2. \tag{3.3}$$

[15]Porównaj Goodman 1961, 1969, Styczeń 1989 .

Zgodnie z proponowanym modelem:

$$\pi^{XY(Z)}_{ab(1)} = 0, \text{ dla dowolnej pary } a, b, \text{ takiej, że } a \neq b \text{ oraz} \quad (3.4)$$

$$\Theta^{XY(Z)}_{ij(2)} = 1, \text{ dla dowolnej pary } i, j. \quad (3.5)$$

Zgodnie z pierwszym warunkiem modelu $MS_0$ osoby niemobilne usytuowane są wyłącznie na głównej przekątnej[16]. Zgodnie z drugim warunkiem wszystkie stosunki szans są równe 1, co opisuje niezależność pomiędzy $X$ i $Y$ wśród osób mobilnych. W języku ruchliwości oznaczałoby to, że dla tej podzbiorowości pozycja społeczno–zawodowa syna nie zależy od pozycji społeczno–zawodowej ojca. Jak wiadomo, warunek ten można również sformułować jako:

$$\pi^{XY(Z)}_{ij(2)} = \pi^{X(Z)}_{i(2)} \pi^{Y(Z)}_{j(2)}.$$

Tak więc w tej podzbiorowości opisywanej za pomocą procesu „mobilności" poszczególne prawdopodobieństwa rozkładu łącznego $X$ oraz $Y$ zależą wyłącznie od częstości brzegowych obydwu tych zmiennych dla $Z = 2$.

Tabela 3.12: Ilustracja modelu $MS_0$ oraz QN – parametry opisujące efekty zmiennych $X$ i $Y$ i interakcję

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $d^X_1 d^Y_1 q_1$ | $d^X_1 d^Y_2$ | $d^X_1 d^Y_3$ | $d^X_1 d^Y_4$ | $d^X_1 d^Y_5$ |
| $x_2$ | $d^X_2 d^Y_1$ | $d^X_2 d^Y_2 q_2$ | $d^X_2 d^Y_3$ | $d^X_2 d^Y_4$ | $d^X_2 d^Y_5$ |
| $x_3$ | $d^X_3 d^Y_1$ | $d^X_3 d^Y_2$ | $d^X_3 d^Y_3 q_3$ | $d^X_3 d^Y_4$ | $d^X_3 d^Y_5$ |
| $x_4$ | $d^X_4 d^Y_1$ | $d^X_4 d^Y_2$ | $d^X_4 d^Y_3$ | $d^X_4 d^Y_4 q_4$ | $d^X_4 d^Y_5$ |
| $x_5$ | $d^X_5 d^Y_1$ | $d^X_5 d^Y_2$ | $d^X_5 d^Y_3$ | $d^X_5 d^Y_4$ | $d^X_5 d^Y_5 q_5$ |

Ilustrację tej hipotezy stanowi tabela 3.12. Należy zauważyć, że w porównaniu do modelu niezależności w modelu $MS_0$ na przekątnej występują parametry interakcji $q_i$ opisujące specyfikę tych komórek. Wartości te określają tendecję do nie–mobilności, tj. określają, ile razy częściej zdarza się występowanie tej samej kategorii dla obydwu zmiennych w stosunku do hipotetycznej sytuacji niezależności opisywanej przez parametry $d^X_i$ oraz $d^Y_j$. Na wielkości tych parametrów wpływ ma odsetek osób niemobilnych i mobilnych ($\beta_1$ i $\beta_2$), jak również odsetki brzegowe zmiennych $X$ i $Y$ w

[16]Możliwe jest sformułowanie innego modelu (Goodman 1972*c*), zgodnie z którym warunek 3.5 obejmuje inne kombinacje zmiennych $X$ i $Y$, tj. nie–mobilność dotyczy innych komórek niż tych położonych na głównej przekątnej. Taki model można również odnieść do tabeli, gdzie kategorie zmiennej wierszowej i kolumnowej są inne.

obydwu pozbiorowościach. Przykładowo, jeśli $\beta_1 = 0$, hipoteza sprowadza się do hipotezy o niezależności, więc $q_i = 1$, dla każdej wartości $i$. Im odsetek $\beta_1$ jest większy, tym wartości $q_i$ sa większe w stosunku do wielkości $d_i^X$ oraz $d_j^Y$.

Warto zauważyć, że komórki na głównej przekątnej definiowane są za pomocą wszystkich parametrów wspomnianych powyżej, co jest odzwierciedleniem wspomnianej powyżej koncepcji Goodmana, zgodnie z którą proces mobilności, nie wyklucza, że pozycja syna będzie taka sama jak pozycja ojca. Parametryzacja tego modelu mogłaby wyglądać następująco:

$$\pi_{ij}^{XY} = d \cdot d_i^X \cdot d_j^Y \cdot q_i \tag{3.6}$$

Przy czym $q_i = 1$, jeśli $i \neq j$. Zgodnie z prezentowaną hipotezą wszystkie parametry $q_i$ są większe bądź równe 1. Jak łatwo zauważyć, gdyby wszystkie parametry $q_i$ były równe 1 model byłby zbieżny z modelem niezależności.

Nieco bardziej ogólny w stosunku do powyższej hipotezy $MS_0$ byłby model, w którym nie będziemy zakładać, że wartości $q_i$ muszą być większe bądź równe 1. Tę postać modelu oznacza się w literaturze jako model quasi–niezależności(QN). Zgodnie z tym modelem może się więc zdarzyć, że dla jednej bądź kilku kategorii obserwujemy tendencję odwrotną do dziedziczenia, tj. pozostanie na tej samej pozycji jest relatywnie mniej prawdopodobne aniżeli wynikałoby to z modelu niezależności. Przy takim ujęciu całą zbiorowość należałoby podzielić na trzy podzbiorowości[17]: tak jak poprzednio należy wyróżnić osoby niemobilne (tj. $Z = 1$), osoby dla których nie ma zależności pomiędzy zmiennymi $X$ i $Y$ (tj. $Z = 2$) oraz nową podzbiorowość ($Z = 3$): dla tych osób, wykluczamy, że obydwie zmienne $X$ i $Y$ przyjmują tę samą wartość, natomiast można mówić o niezależności o jeśli chodzi o pozostałe kombinacje obydwu zmiennych. Formułując model bardziej precyzyjnie, proporcje poszczególnych podzbiorowości wyróżnionych ze względu na zmienną Z wynoszą odpowiednio $\beta_1$, $\beta_2$ oraz $\beta_3$ i sumują się do 1. Prawdopodobieństwa rozkładu łącznego zmiennych $X$ i $Y$ wynoszą więc:

$$\pi_{ij}^{XY} = \pi_{ij(1)}^{XY(Z)}\beta_1 + \pi_{ij(2)}^{XY(Z)}\beta_2 + \pi_{ij(3)}^{XY(Z)}\beta_3. \tag{3.7}$$

Do warunków 3.4 oraz 3.5, dochodzi następujące założenie dotyczące ostatniej z wymienionych podzbiorowości:

$$\pi_{ii(3)}^{XY(Z)} = 0 \text{ oraz } \Theta_{a/b;c/d;(3)}^{X\ \ Y\ \ (Z)} = 1 \tag{3.8}$$

dla dowolnej wartości $i$, jak również dowolnej pary wartości zmiennej $X$ i zmiennej $Y$, takich, że $a \neq c$, $a \neq d$, $b \neq c$, $b \neq d$. Warunek powyższy określa, że komórki na

[17] Inaczej model ten wyjaśnia Goodman (1969), definiując tzw. model dwu–etapowy (two–stage model), zgodnie z którym część osób „odchodzi" z głównej przekątnej.

głównej przekątnej są równe 0, a wszystkie stosunki szans (nie tylko lokalne), które nie obejmują komórek głównej przekątnej są równe 1.

Tabela 3.13: Ilustracja modelu quasi-niezależności QN — rozkłady prawdopodobieństwa w trzech podzbiorowościach wyróżnionych ze względu na zmienną $Z$

| $Z=1$ | | | | | |
|---|---|---|---|---|---|
| $X\backslash Y$ | 1 | 2 | 3 | 4 | 5 |
| 1 | $w_1$ | 0 | 0 | 0 | 0 |
| 2 | 0 | $w_2$ | 0 | 0 | 0 |
| 3 | 0 | 0 | $w_3$ | 0 | 0 |
| 4 | 0 | 0 | 0 | $w_4$ | 0 |
| 5 | 0 | 0 | 0 | 0 | $w_5$ |
| $Z=2$ | | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | 5 |
| 1 | $a_1^X a_1^Y$ | $a_1^X a_2^Y$ | $a_1^X a_3^Y$ | $a_1^X a_4^Y$ | $a_1^X a_5^Y$ |
| 2 | $a_2^X a_1^Y$ | $a_2^X a_2^Y$ | $a_2^X a_3^Y$ | $a_2^X a_4^Y$ | $a_2^X a_5^Y$ |
| 3 | $a_3^X a_1^Y$ | $a_3^X a_2^Y$ | $a_3^X a_3^Y$ | $a_3^X a_4^Y$ | $a_3^X a_5^Y$ |
| 4 | $a_4^X a_1^Y$ | $a_4^X a_2^Y$ | $a_4^X a_3^Y$ | $a_4^X a_4^Y$ | $a_4^X a_5^Y$ |
| 5 | $a_5^X a_1^Y$ | $a_5^X a_2^Y$ | $a_5^X a_3^Y$ | $a_5^X a_4^Y$ | $a_5^X a_5^Y$ |
| $Z=3$ | | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | $b_1^X b_2^Y$ | $b_1^X b_3^Y$ | $b_1^X b_4^Y$ | $b_1^X b_5^Y$ |
| 2 | $b_2^X b_1^Y$ | 0 | $b_2^X b_3^Y$ | $b_2^X b_4^Y$ | $b_2^X b_5^Y$ |
| 3 | $b_3^X b_1^Y$ | $b_3^X b_2^Y$ | 0 | $b_3^X b_4^Y$ | $b_3^X b_5^Y$ |
| 4 | $b_4^X b_1^Y$ | $b_4^X b_2^Y$ | $b_4^X b_3^Y$ | 0 | $b_4^X b_5^Y$ |
| 5 | $b_5^X b_1^Y$ | $b_5^X b_2^Y$ | $b_5^X b_3^Y$ | $b_5^X b_4^Y$ | 0 |

Wszystkie trzy podzbiorowości ilustruje tabela 3.13. Przypuśćmy, że mamy do czynienia z tablicą ruchliwości edukacyjnej i kategoriami 1. podstawowe, 2. niepełne średnie, 3. ukończone średnie, 4. niepełne wyższe, 5. ukończone wyższe. Pierwsza z nich ilustruje odsetki osób niemobilnych, przyjmujących tę samą kategorię obydwu zmiennych. Przykładowo wielkość $w_3$ pokazuje odsetek osób niemobilnych, które „odziedziczyły ” wykształcenie średnie po swoim ojcu. Kolejne tabele ilustrują osoby mobilne, dla których nie wykluczamy ($Z = 2$) bądź wykluczamy przyjmowanie tej samej pozycji ($Z = 3$). W pierwszej z nich mamy do czynienia z kompletną niezależnością wykształcenia ojca od wykształcenia syna, w drugiej identyczność warunkowych

szans nie obejmuje komórek na głównej przekątnej. Przykładowo dla obydwu tych podzbiorowości ($Z = 2$ lub $Z = 3$), szanse na to, że zmienna $Y$ przyjmie raczej wartość $y_2$ aniżeli $y_3$, nie zależy od tego, czy mamy do czynienia z wartością $x_1$, czy wartością $x_5$. Tak więc proporcja osób, które mają wykształcenie niepełne średnie do osób, które mają wykształcenie ukończone średnie jest taka sama wśród osób, których ojcowie mają wykształcenie podstawowe jak wśród osób, których ojcowie mają wykształcenie wyższe. Warto również zauważyć, że wyróżniona szansa wynosi $a_2^Y/a_3^Y$ dla podzbiorowości $Z = 2$, natomiast dla podzbiorowości $Z = 3$ wynosi $b_2^Y/b_3^Y$, czyli nie muszą one wynosić tyle samo. Ponadto dla podzbiorowości $Z = 2$ równe 1 są również stosunki szans, które obejmują komórki związane z główną przekątną, tj. wyróżniona powyżej szansa była tak sama gdybyśmy porównywali ojców o wykształceniu niepełnym średnim lub średnim.

Prezentowana wcześniej tabela 3.12, która stanowiła ilustrację modelu $MS_0$ może być również interpretowana w kontekście ogólniejszego modelu quasi–niezależności QN. Tak jak zostało zasygnalizowane powyżej wybrane parametry $q_i$ mogą być mniejsze od 1. Na wielkości parametrów $d_i^X$, $d_j^Y$, $q_i$ mają wpływ zarówno wielkości z tabeli 3.13, tj. $a_i^X$, $a_j^Y$, $b_i^X$, $b_j^Y$, $w_i$, jak też proporcje pomiędzy grupami tj. wielkości $\beta_1$, $\beta_2$ i $\beta_3$. Przykładowo, jeśli nie ma osób niemobilnych ($\beta_1 = 0$), to mamy do czynienia z sytuacją, gdy wszystkie parametry $q_i$ są mniejsze od 1.

Warto w tym miejscu dodać, że model powyższy nie odpowiada na pytanie o odsetki poszczególnych podzbiorowości wyróżnionych ze względu na zmienną $Z$, nie można też zrekonstruować rozkładów prawdopodobieństwa w poszczególnych podzbiorowościach opisanych powyżej. Ujmując to inaczej: ten sam rozkład w całej zbiorowości zgodny z hipotezą o quasi–niezależności można „zdekomponować" na różne rozkłady w podzbiorowościach wyróżnionych ze względu na wartości zmiennej $Z$. Co więcej, podzbiorowości te mogą mieć różne odsetki, choć oczywiście relacje pomiędzy tymi odsetkami nie są dowolne. Przykład pokazujący dwie różne dekompozycje tablicy o wymiarach 4 x 4 został przedstawiony w Aneksie.

Hipoteza o quasi–niezależności posiada dodatkowo $r$ niezależne parametry $q_i$ w stosunku do hipotezy o niezależności. Liczba stopni swobody wynosi $df = (r-1)^2 - r$. Hipotezę tę można sformułować również bez wprowadzania zmiennej $Z$. Wystarcza jeśli w odniesieniu do rozkładu łącznego zmiennych $X$ oraz $Y$, założymy, że wszystkie stosunki szans są symetryczne, a te które nie obejmują komórek głównej przekątnej są równe 1. Wydaje się jednak, że posłużenie się trzema podzbiorowościami w bardziej perswazyjny sposób tłumaczy proces jaki opisuje ta hipoteza[18]. Możliwe jest również

[18]Sformułowanie hipotezy za pomocą stosunków szans, jest bardziej ogólne niż formuły 3.4,3.4, 3.8, gdyż formuła 3.8, nie ma zastosowania w odniesieniu do tabeli o wymiarach 3 x 3. Jak zobaczymy

sformułowanie hipotezy o quasi–niezależności wykorzystując jedynie lokalne stosunki szans. Hipoteza quasi–niezależności w tej postaci zamieszczona została w Aneksie, można tam również znaleźć bardziej precyzyjne wyjaśnienie liczby stopni swobody dla tego modelu.

Tabela 3.14: Rozkład oczekiwany dla danych z tabeli 3.6 zgodny z modelem QN

| Czy przy aktualnym dochodzie netto P Pana(i) gospodarstwo domowe wiąże koniec z końcem? | | | | | | |
|---|---|---|---|---|---|---|
| Odpowiedzi w 2000 roku ($X$) | Odpowiedzi w 2005 roku ($Y$) | | | | | Suma |
| | 1 | 2 | 3 | 4 | 5 | |
| 1. Z wielką trudnością | 271,7 | 118,9 | 131,9 | 49,0 | 10,7 | 582 |
| 2. Z trudnością | 71,0 | 134,5 | 172,7 | 64,2 | 14,0 | 456 |
| 3. Z pewną trudnością | 69,3 | 151,9 | 248,5 | 62,6 | 13,7 | 546 |
| 4. Raczej łatwo | 24,1 | 52,8 | 58,5 | 73,9 | 4,7 | 214 |
| 5. Łatwo | 3,6 | 8,0 | 8,8 | 3,3 | 13,5 | 37 |
| Suma | 440 | 466 | 620 | 253 | 57 | 1836 |

W tabeli 3.14 przedstawione zostały liczebności oczekiwane zgodne z modelem quasi–niezależności dla danych panelowych. Tabela ta pokazuje, że informacje z próby potrzebne do oszacowania rozkładu oczekiwanego metodą największej wiarygodności dotyczą komórek znajdujących się na głównej przekątnej, i — podobnie jak w przypadku hipotezy o niezależności — rozkładów brzegowych, tj. zakłada się, że:

$$\widehat{\pi}_i^X = p_i^X, \quad \widehat{\pi}_j^Y = p_j^Y, \quad \widehat{\pi}_{ii}^{XY} = p_{ii}^{XY}.$$

Warto dodać, że w estymacji prawdopodobieństw znajdujących się poza główną przekątną nie uwzględnia się komórek głównej przekątnej. Mówiąc inaczej nie bierze się ich pod uwagę dla wyznaczenia rozkładu w podzbiorowości osób mobilnych. W tym sensie te ostatnie opisuje się często jako „pominięte" (omitted) bądź „usunięte" (deleted, blanked out). Zwrot ten jest często używany w literaturze (np. Goodman 1969, 2007), jeśli chce się zasygnalizować, że w modelu uwzględniamy specyfikę komórek na głównej przekątnej.

Analizując rozkład oczekiwany, można również zauważyć, że wszystkie stosunki szans (niekoniecznie lokalne), nie zawierające komórek związanych z główną przekątną

w dalszej części, dla tabeli o takich wymiarach hipoteza o quasi–niezależności jest równoważna hipotezie o quasi-symetrii, w tym sensie formuła, która obejmuje tabele o wymiarach 4 x 4 lub większe wydaje się wystarczająco ogólna.

są równe 1. Przykładowo, dla tabeli 3.14 daje się na przykład pokazać, że:

$$\Theta_{13}^{XY} = \frac{131,9 \cdot 64,2}{49,0 \cdot 172,7} = 1,$$

bądź

$$\Theta_{2/4;1/5}^{X\ Y} = \frac{71,0 \cdot 4,7}{14,0 \cdot 24,1} = 1.$$

Model quasi–niezależności w literaturze na ogół definiuje się za pomocą równania 3.6, tj. posługując się parametrami modelu, nie zaś za pomocą stosunków szans. W tej pracy niejednokrotnie było podkreślone, że każdą hipotezę, można parametryzować na wiele sposobów, stąd też o istocie modelu quasi-niezależności stanowią raczej warunki 3.4-3.8. Warto zwrócić uwagę na tabelę 3.15, która stanowi przykład parametryzacji tego samego modelu. Wielkości $a$, $q_i$ można interpretować jako parametry interakcji modelu, który można wyrazić jako:

$$\pi_{ij}^{XY} = \begin{cases} d \cdot d_i^X \cdot d_j^Y & \text{gdy } i{=}1 \text{ lub } j{=}1 \\ d \cdot d_i^X \cdot d_j^Y \cdot q_i & \text{gdy } i = j > 1 \\ d \cdot d_i^X \cdot d_j^Y \cdot a & \text{w pozostałych przypadkach.} \end{cases} \tag{3.9}$$

Model ten na pierwszy rzut oka wygląda zupełnie inaczej niż model quasi–niezależności przedstawiony w tabeli 3.12. Okazuje się, jednak, że modele te są tożsame. Przykład ten pokazuje raz jeszcze, że każdą hipotezę można parametryzować na wiele konkurencyjnych sposobów, a wybór konkretnej z nich ma ułatwiać jego interpretację, jak również pozwalać na pomiar interesujących nas interakcji. Na przykład, w odniesieniu do omawianego modelu, parametry głównej przekątnej $q_i$ z równania 3.6 wskazują na tendencję do dziedziczenia pozycji.

Tabela 3.15: Ilustracja alternatywna modelu quasi–niezależności QN — parametry opisujące efekty główne zmiennych $X$ oraz $Y$ i interakcję między nimi

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $d_1^X d_1^Y$ | $d_1^X d_2^Y$ | $d_1^X d_3^Y$ | $d_1^X d_4^Y$ | $d_1^X d_5^Y$ |
| $x_2$ | $d_2^X d_1^Y$ | $d_2^X d_2^Y q_2 1$ | $d_2^X d_3^Y a$ | $d_2^X d_4^Y a$ | $d_2^X d_5^Y a$ |
| $x_3$ | $d_3^X d_1^Y$ | $d_3^X d_2^Y a$ | $d_3^X d_3^Y q_3$ | $d_3^X d_4^Y a$ | $d_3^X d_5^Y a$ |
| $x_4$ | $d_4^X d_1^Y$ | $d_4^X d_2^Y a$ | $d_4^X d_3^Y a$ | $d_4^X d_4^Y q_4$ | $d_4^X d_5^Y a$ |
| $x_5$ | $d_5^X d_1^Y$ | $d_5^X d_2^Y a$ | $d_5^X d_3^Y a$ | $d_5^X d_4^Y a$ | $d_5^X d_5^Y q_5$ |

Jak pokazuje tabela 3.11, dopasowanie modelu quasi-niezależności do danych z analizowanych tabel 3.2, 3.5, 3.6 jest znacznie lepsze niż modeli N i NS, niemniej

nie jest ono zadowalające. W przypadku ruchliwości zawodowej redukcja statystyki statystyki $G^2$ w stosunku do modelu niezależności wynosi $G^2 = 432,0 - 104,2 = 327,8$ przy czym obydwa modele różni 6 stopni swobody. Analogicznie test warunkowy dla ruchliwości edukacyjnej wynosi $G^2 = 337,8 - 105,1 = 232,7$, $df = 4$ a dla danych panelowych $G^2 = 527,8 - 185,4 = 342,4$, $df = 5$. W każdym przypadku redukcja jest istotna statystycznie, przy czym pamiętać należy, że stosowanie testów warunkowych jest uprawnione, gdy zakładamy, że bardziej złożony z dwóch modeli zagnieżdżonych (w naszym przypadku QN) jest prawdziwy, co nie wydaje się uprawnione w świetle wyników weryfikacji. Niemniej powyższe testy warunkowe wyraźnie sugerują, że przy formułowaniu hipotez pożądane może być uwzględnienie specyfiki głównej przekątnej.

Parametry tych komórek wynoszą dla tabeli ruchliwości edukacyjnej odpowiednio: $q_1 = 6,46$, $q_2 = 0,71$, $q_3 = 1,42$, $q_4 = 9,69$. Wielkości te wskazują na przykład, że dziedziczenie wykształcenia wyższego zdarza się 6,46 razy częściej, aniżeli w hipotetycznej sytuacji niezależności, tj. braku barier społecznych w uzyskiwaniu wykształcenia. Zauważyć należy, że wszystkie parametry głównej przekątnej są większe od 1, poza parametrem wyznaczonym dla wykształcenia średniego, przy czym najsilniejsza jest tendencja do dziedziczenia wykształcenia podstawowego i niepełnego podstawowego. Analogiczne parametry dla danych panelowych są równe $q_1 = 5,01$, $q_2 = 0,86$, $q_3 = 1,48$, $q_4 = 3,39$, $q_5 = 18,86$. Pokazuje to, że z wyjątkiem drugiej kategorii, respondenci wskazują relatywnie częściej na kombinacje opisujące te same odpowiedzi w 2000 i 2005. Można również analizować parametry opisujące rozkłady obydwu zmiennych. Przykładowo: $d_1^Y = 1,01$, $d_5^Y = 0,2$, co pokazuje, że szansa na to, że respondent w 2005 roku wskazał raczej, że radził sobie „z wielką trudnością" aniżeli że radził sobie „łatwo" wynosi w przybliżeniu $d_1^Y/d_5^Y \approx 5$ dla trzech podzbiorowości: osób, które w 2000 roku udzieliły odpowiedzi „z trudnością", „z pewną trudnością" lub „raczej łatwo". Wynika to z tego, że zgodnie z modelem QN każdy stosunek szans nie obejmujący komórek głównej przekątnej wynosi 1. Warto jednak podkreślić raz jeszcze, że model quasi–niezależności należałoby odrzucić na standardowo przyjmowanych poziomach istotności, tak więc wyciąganie wniosków na podstawie tego modelu i jego parametrów jest kontrowersyjne, w tym sensie, że jest bardzo prawdopodobne, że nie odzwierciedlają one faktycznych zależności.

Powyżej wyróżnione zostały trzy hipotezy: symetrycznej niezależności (NS), niezależności (N) i quasi–niezależności (QN). Jak łatwo zauważyć pierwsza z nich jest hipotezą najprostszą, gdyż poza założeniem o niezależności zawiera dodatkowo postulat identyczności rozkładów brzegowych zmiennej wierszowej i kolumnowej, natomiast trzecia hipoteza jest najbardziej złożona, ponieważ zmienne mogą być w pewnym zakresie zależne, co jest wynikiem uwzględnienia specyfiki komórek przekątnej.

Warto zwrócić uwagę na modyfikacje, które pozwalają na sformułowanie dodatkowych modeli. W modelu quasi–niezależności parametry dotyczące głównej przekątnej są specyficzne, tj. model ten uwzględnia tendencję do „dziedziczenia" pozycji, wskazywania na tę samą odpowiedź w dwóch porównywanych punktach czasowych, itd. Można uprościć model QN, zakładając dodatkowo, że tendencja ta jest taka sama dla każdej kategorii analizowanych zmiennych. Model taki oznaczany będzie jako QhN. Ilustrację tego modelu stanowi tablica 3.16.

Tabela 3.16: Ilustracja modelu quasi-niezależności QhN - parametry opisujące efekty główne zmiennych $X$ i $Y$ i parametr przekątnej

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $d_1^X d_1^Y q$ | $d_1^X d_2^Y$ | $d_1^X d_3^Y$ | $d_1^X d_4^Y$ | $d_1^X d_5^Y$ |
| $x_2$ | $d_2^X d_1^Y$ | $d_2^X d_2^Y q$ | $d_2^X d_3^Y$ | $d_2^X d_4^Y$ | $d_2^X d_5^Y$ |
| $x_3$ | $d_3^X d_1^Y$ | $d_3^X d_2^Y$ | $d_3^X d_3^Y q$ | $d_3^X d_4^Y$ | $d_3^X d_5^Y$ |
| $x_4$ | $d_4^X d_1^Y$ | $d_4^X d_2^Y$ | $d_4^X d_3^Y$ | $d_4^X d_4^Y q$ | $d_4^X d_5^Y$ |
| $x_5$ | $d_5^X d_1^Y$ | $d_5^X d_2^Y$ | $d_5^X d_3^Y$ | $d_5^X d_4^Y$ | $d_5^X d_5^Y q$ |

Hipotezę tę można sformułować bardziej precyzyjnie. Do warunków 3.4–3.8, należy dodać następujące założenia:

$$\frac{\pi_{ii(1)}^{XY(Z)}}{\pi_{jj(1)}^{XY(Z)}} = \frac{\pi_{ii(2)}^{XY(Z)}}{\pi_{jj(2)}^{XY(Z)}}, \text{ dla dowolnych wartości i, j oraz} \tag{3.10}$$

$$\frac{\pi_{ab(3)}^{XY(Z)}}{\pi_{cd(3)}^{XY(Z)}} = \frac{\pi_{ab(2)}^{XY(Z)}}{\pi_{cd(2)}^{XY(Z)}}, \text{ dla dowolnych wartości } a \neq b,\ c \neq d. \tag{3.11}$$

Warunek 3.10 głosi, że proporcje pomiędzy osobami niemobilnymi dla poszczególnych komórek jest taka sama jak analogiczna proporcja wyznaczona dla podziorowości $Z = 2$. Warunek 3.10 głosi, że iloraz dwóch dowolnych prawdopodobieństw dotyczących komórek ulokowanych poza przekątną, będzie taki sam dla pozbiorowości $Z = 2$ oraz $Z = 3$. W stosunku do ilustracji z tabeli 3.13 model QhN, zakłada, że $w_i = a_i^X a_i^Y$, oraz $b_i^X = a_i^X$ jak również $b_j^Y = a_j^Y$. Jeżeli $\beta_1 = 0$, czyli nie ma osób niemobilnych, wówczas parametr $q$ w tabeli 3.16 jest mniejszy od 1, jeśli $\beta_3 = 0$ wówczas $q > 1$. Jeśli spełnione są obydwa warunki, wówczas mamy do czynienia z sytuacją niezależności. Model powyższy podobnie jak model QN nie przesądza o wielkości poszczególnych podzbiorowości wyznaczonych ze względu na zmienną $Z$. Ten sam rozkład zgodny z hipotezą QhN można zdekomponować zakładając różne odsetki

poszczególnych podzbiorowości, tj. różne wielkości $\beta_1$, $\beta_2$, $\beta_3$. Oczywiście wielkości te nie mogą być dowolne a pozostają w pewnych relacjach[19].

Daje się pokazać, że parametr $q$ zdaje sprawę z relacjami pomiędzy odsetkiem osób ulokowanych na głównej przekątnej a pozostałymi osobami, tj:

$$q = \frac{\sum_{i=1}^{r} \pi_{ii}^{XY} / \sum_{i=1}^{r} \sum_{j=1, i \neq j}^{r} \pi_{ij}^{XY}}{\sum_{i=1}^{r} \pi_{ii(2)}^{XY(Z)} / \sum_{i=1}^{r} \sum_{j=1, i \neq j}^{r} \pi_{ij(2)}^{XY(Z)}}. \tag{3.12}$$

Mianownik tego wyrażenia wskazuje na stosunek prawdopodobieństwa, że mamy do czynienia z tą samą wartością zmiennej wierszowej i kolumnowej do prawdopodobieństwa, że zmienne mają inne wartości, przy czym wielkość ta jest wyznaczona dla podzbiorowości $Z = 2$, w której zmienne są niezależne. Licznik wskazuje na ten analogiczny stosunek wyznaczony dla rozkładu zgodnego z modelem QhN dla całej zbiorowości, tj. obejmującej wszystkie podzbiorowości wyróżnione ze względu na zmienną $Z$. Parametr $q$ wskazuje więc, ile razy — zgodnie z modelem — częściej (rzadziej) zdarzają się osoby „dziedziczące" w stosunku do pozostałych osób aniżeli w sytuacji niezależności.

Model ten można również zdefiniować wyłącznie za pomocą lokalnych stosunków szans bez wyodrębniania podzbiorowości ze względu na zmienną $Z$, co zostało przedstawione w Aneksie. Jeśli chodzi o parametryzację tego modelu, w stosunku do modelu QN (3.6) można dodatkowo założyć, że $q_i = q$, dla każdej kategorii $i$:

$$\pi_{ij}^{XY} = d \cdot d_i^X \cdot d_j^Y \cdot q, \tag{3.13}$$

przy czym $q = 1$, jeśli $i \neq j$. W modelu tym nie wykluczamy tendencji do współwystępowania tych samych kategorii obydwu zmiennych, przy czym jest ona stała dla każdej komórki na głównej przekątnej. Jest to więc model pośredni pomiędzy modelem niezależności i quasi–niezależności. W stosunku do niezależności stochastycznej model ten posiada tylko jeden parametr więcej, dlatego jego liczba stopni swobody wynosi $df = (r-1)^2 - 1$

W tabeli 3.17 podane zostały liczebności oczekiwane zgodne z tym modelem dla tablicy ruchliwości edukacyjnej. Do oszacowania rozkładu oczekiwanego metodą największej wiarygodności potrzebne są informacje o rozkładach brzegowych i sumie prawdopodobieństw komórek na głównej przekątnej, tj. zakłada się, że:

$$\widehat{\pi}_i^X = p_i^X, \quad \widehat{\pi}_j^Y = p_j^Y, \quad \sum_i^r \widehat{\pi}_{ii}^{XY} = \sum_i^r p_{ii}^{XY}.$$

[19]Gdy $\beta_2 = 0$, gdyż wówczas trudno zastosować warunki 3.10, 3.11. Aby były one adekwatne należy przyjąć, że $\beta_2 > 0$.

Tabela 3.17: Rozkład oczekiwany dla danych z tablicy 3.5 zgodny z modelem QhN

| Wykształcenie ojca | Wykształcenie syna | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | Suma |
| 1. Podstawowe i niepełne podstawowe | 313,5 | 430,7 | 224,2 | 176,7 | 1145,0 |
| 2. Niepełne średnie (w tym zasadnicze zawodowe) | 28,1 | 163,0 | 41,3 | 32,6 | 265,0 |
| 3. Ukończone średnie | 23,5 | 66,3 | 70,9 | 27,2 | 188,0 |
| 4. Niepełne wyższe i wyższe | 9,9 | 28,0 | 14,6 | 23,6 | 76,0 |
| Suma | 375,0 | 688,0 | 351,0 | 260,0 | 1674,0 |

Tabela 3.11 pokazuje, że model QhN jest słabo dopasowany. Parametr q wynosi odpowiednio 2,05 dla tablicy ruchliwości edukacyjnej i 2,06 dla danych panelowych. I tym razem należy pamiętać o zastrzeżeniu, że modele te są nierealistyczne, więc wartości powyższych parametrów nie muszą opisywać faktycznej skłonności do dziedziczenia pozycji i udzielenia tej samej odpowiedzi w 2000 i 2005 roku.

Możliwe jest formułowanie kolejnych modeli, na przykład, dodanie do każdego z rozpatrywanych modeli quasi–niezależności QN, QhN dodatkowego założenia o identyczności rozkładów brzegowych. Daje się również sformułować model pośredni pomiędzy modelami QN oraz QhN poprzez dodanie do modelu QN założenia o równości jedynie wybranych parametrów interakcji przekątnej, np: $q_1 = q_r$. Modele te nie będą szczegółowo omawiane, ponieważ nawet model quasi–niezależności (najbardziej złożony z prezentowanych do tej pory) rzadko jest realistycznym opisem rzeczywistości. Sygnalizujemy jednak taką możliwość, aby podkreślić wielość hipotez jakie można sformułować w odniesieniu do tablic o takich samych kategoriach zmiennej wierszowej i zmiennej kolumnowej. Jak zobaczymy w dalszej części tego rozdziału, uchylenie założenia o quasi–niezależności — tj. identyczności szans, które nie dotyczą głównej przekątnej — pozwoli na sformułowanie kolejnych hipotez.

### 3.2.2 Model symetrii i quasi–symetrii

Do tej pory zakładaliśmy, że zależność pomiędzy zmiennymi dotyczy tylko i wyłącznie komórek na przekątnej. W praktyce założenie to okazuje się zbyt restrykcyjne, gdyż zależność pomiędzy zmiennymi w tablicach ruchliwości nie dotyczy jedynie dziedziczenia i niezmienności pewnych cech w czasie. Przykładowo, można oczekiwać, że

osoby, których ojcowie mieli wykształcenie wyższe częściej będą mieli wykształcenie niepełne średnie niż podstawowe w porównaniu do osób, których ojcowie mieli wykształcenie ukończone średnie. W kolejnych modelach związek pomiędzy zmiennymi nie będzie dotyczyć wyłącznie komórek na przekątnej, niemniej będzie się zakładać, że związek ten jest symetryczny. Odnośnie do powyższego przykładu, jeśli opisany związek jest symetryczny oznacza to, że podobna relacja występuje jeśli porównuje się osoby, których ojcowie mieli wykształcenie niepełne średnie i podstawowe. Oczekiwać będziemy, że wśród tych pierwszych proporcja liczby osób z wykształceniem wyższym do liczby osób z wykształceniem średnim będzie większa niż wśród tych drugich.

Poniżej omówione zostaną modele symetrii i quasi–symetrii. Intuicja zawarta w powyższym akapicie wymaga bardziej precyzyjnego sformułowania. Model symetrii — oznaczany jako S — można zdefiniować odwołując się wyłącznie do prawdopodobieństw rozkładu łącznego. Hipoteza z nim związana głosi, że prawdopodobieństwa dla komórek symetrycznie położonych wokół przekątnej są identyczne tj.

$$\pi_{ij}^{XY} = \pi_{ji}^{XY}, \tag{3.14}$$

dla dowolnej pary wartości $i$, $j$. W odniesieniu do tablicy ruchliwości oznaczałoby to między innymi, że odsetek osób z wykształceniem wyższym, których ojciec miał wykształcenie podstawowe jest identyczny jak odsetek badanych z wykształceniem podstawowym, których ojciec miał wykształcenie wyższe. Tabela 3.18 ilustruje hipotezę o symetrii.

Tabela 3.18: Rozkład łączny ilustrujący model symetrii

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $\pi_{11}$ | $\pi_{12}$ | $\pi_{13}$ | $\pi_{14}$ | $\pi_{15}$ |
| $x_2$ | $\pi_{12}$ | $\pi_{22}$ | $\pi_{23}$ | $\pi_{24}$ | $\pi_{25}$ |
| $x_3$ | $\pi_{13}$ | $\pi_{23}$ | $\pi_{33}$ | $\pi_{34}$ | $\pi_{35}$ |
| $x_4$ | $\pi_{14}$ | $\pi_{24}$ | $\pi_{34}$ | $\pi_{44}$ | $\pi_{45}$ |
| $x_5$ | $\pi_{15}$ | $\pi_{25}$ | $\pi_{35}$ | $\pi_{45}$ | $\pi_{55}$ |

Zauważmy, że warunek 3.14 implikuje identyczność rozkładów brzegowych obydwu zmiennych. Hipotezę tę można zdefiniować jako model logarytmiczno–liniowy. Należy założyć, że parametry opisujące efekty obydwu zmiennych dla kolejnych kategorii są sobie równe, a z drugiej strony parametry interakcji są symetryczne względem

przekątnej[20], tj:

$$d_i^X = d_i^Y \text{ oraz} \tag{3.15}$$

$$d_{ij}^{XY} = d_{ji}^{XY} \text{ dla każdej pary } i, j. \tag{3.16}$$

Model ten posiada 1 parametr $d$, efekty zmiennych $X$ i $Y$ opisuje łącznie $r-1$ niezależnych parametrów $d_i^X$, interakcję opisuje $(r-1)r/2$ parametrów $d_{ij}^{XY}$. Liczba stopni swobody dla modelu symetrii wynosi więc:

$$df = r^2 - (r-1)r/2 - r = (r-1)r/2.$$

Jak łatwo zauważyć, jest to zbieżne z liczbą założeń wynikających z warunku 3.14. Równe sobie są prawdopodobieństwa na kolejnych pseudo–przekątnych a liczba tych prawdopodobieństw wynosi $(r-1)r/2$. Liczba tych warunków jest tożsama z liczbą stopni swobody[21]:

$$df = 1 + 2 + \ldots + (r-1) = (r-1)r/2.$$

W odniesieniu do naszych danych z tablic 3.2, 3.5, 3.6 model S należałoby odrzucić przy standardowo przyjmowanych poziomach istotności, co pokazuje tabela 3.11. Stosunkowo najlepiej wydaje się on dopasowany do danych panelowych $G^2 = 61,6$, $df = 10$ ($p < 0,0001$), a indeks rozbieżności wskazuje, że niezgodnie z tym modelem jest zaklasyfikowanych około 6% gospodarstw w próbie. Niemniej, model ten niej jest akceptowalny na standardowo przyjmowanych poziomach istotności.

Rozkład oczekiwany zgodny z symetrią dla danych panelowych przedstawia tabela 3.19. Jak widać symetryczne są wszystkie liczebności — i co się z tym wiąże prawdopodobieństwa — w konsekwencji identyczne są rozkłady brzegowe obydwu zmiennych, tj. w porównywanych punktach czasowych nie zmienił się rozkład zmiennej opisującej sytuację materialną gospodarstwa. Porównując ten rozkład z danymi empirycznymi można zauważyć, że poszczególne liczebności oczekiwane (prawdopodobieństwa) są średnią z dwóch liczebności empirycznych (prawdopodobieństw) z komórek położonych symetrycznie względem głównej przekątnej, tj. zgodnie z metodą największej wiarygodności zachodzi:

$$\widehat{\pi}_{ij}^{XY} = \frac{p_{ij}^{XY} + p_{ji}^{XY}}{2}.$$

[20] Zakładamy, że przyjęta parametryzacja nie jest „asymetryczna". Przykładowo, jeśli przyjmiemy różne kategorie odniesienia poszczególnych zmiennych, pomimo istnienia symetrii parametry $d_{ij}^{XY}$ oraz $d_{ji}^{XY}$ nie byłyby sobie równe, podobnie nie byłyby równe efekty poszczególnych zmiennych dla tej samej kategorii. Jeśli jednak przyjmiemy jako kategorię odniesienia tę samą kategorię dla obydwu zmiennych bądź powszechnie stosowaną parametryzację odchyleń multiplikatywnych, obydwa warunki będą spełnione.

[21] W tej i kilku kolejnych formułach dotyczących liczby stopni swobody będziemy posługiwać się formułą na sumę skończonego ciągu arytmetycznego, tj. $1 + 2 + \ldots + n = (n+1)(n/2)$.

Tabela 3.19: Rozkład oczekiwany zgodny z modelem symetrii dla danych z tabeli 3.6

| Czy przy aktualnym dochodzie netto Pana(i) gospodarstwo domowe wiąże koniec z końcem? | | | | | | |
|---|---|---|---|---|---|---|
| Odpowiedzi w 2000 roku ($X$) | Odpowiedzi w 2005 roku ($Y$) | | | | | Suma |
| | 1 | 2 | 3 | 4 | 5 | |
| 1. Z wielką trudnością | 271,7 | 135,6 | 85,2 | 16,4 | 2,1 | 511,0 |
| 2. Z trudnością | 135,6 | 134,5 | 150,5 | 36,1 | 4,5 | 461,2 |
| 3. Z pewną trudnością | 85,2 | 150,5 | 248,5 | 89,5 | 9,3 | 583,1 |
| 4. Raczej łatwo | 16,4 | 36,1 | 89,5 | 73,9 | 17,6 | 233,5 |
| 5. Łatwo | 2,1 | 4,5 | 9,3 | 17,6 | 13,4 | 46,9 |
| Suma | 511,0 | 461,2 | 583,1 | 233,5 | 46,9 | 1835,8 |

Na ogół bardziej realistyczny — szczególnie w odniesieniu do ruchliwości — okazuje model quasi–symetrii (QS). Nie zakłada się w nim nic na temat identyczności rozkładów brzegowych a symetria dotyczy jedynie związku pomiędzy zmiennymi. Z warunków 3.15 oraz 3.16, które dotyczyły modelu symetrii, w modelu quasi-symetrii musi być spełniony jedynie drugi z nich. Formułując ten model za pomocą lokalnych stosunków szans spełniony jest warunek:

$$\Theta_{ij}^{XY} = \Theta_{ji}^{XY} \tag{3.17}$$

dla każdej pary $i$, $j$, takich, że $i < r-1$, $j \leq r-1$, oraz $i < j$. Z powyższego zapisu wynika, że każdy stosunek szans wyodrębniony dla dowolnych kategorii jednej i drugiej zmiennej jest symetryczny. Model ten w postaci parametrycznej można przedstawić jako:

$$\pi_{ij}^{XY} = d \cdot d_i^X \cdot d_j^Y \cdot s_{ij}^{XY}, \tag{3.18}$$

gdzie $s_{ij}^{XY}$ jest parametrem opisującym interakcję dwóch zmiennych i zachodzi $s_{ij}^{XY} = s_{ji}^{XY}$, przy założeniu, że mamy do czynienia z parametryzacją względem tej samej kategorii odniesienia $x_i$, $y_i$ [22]. Przykładowo w ilustracji w tabeli 3.20 jako kategorie odniesienia zostały przyjęte pierwsze kategorie obydwu zmiennych.

Model ten posiada $r(r-1)/2$ niezależnych parametrów interakcji więcej niż model niezależności stochastycznej, tak więc posiada $df = (r-1)(r-2)/2$ stopni swobody.

[22]Tak jak było sygnalizowane, w odniesieniu do formuł 3.15, 3.15, możliwe byłoby przyjęcie parametryzacji, w której pomimo istnienia quasi–symetrii parametry $s_{ij}^{XY}$ oraz $s_{ji}^{XY}$ nie byłyby sobie równe, na przykład gdybyśmy przyjęli inne kategorie odniesienia dla zmiennej wierszowej i kolumnowej.

Tabela 3.20: Ilustracja modelu quasi-symetrii — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $s_{22}$ | $s_{23}$ | $s_{24}$ | $s_{25}$ |
| $x_3$ | 1 | $s_{32}$ | $s_{33}$ | $s_{34}$ | $s_{35}$ |
| $x_4$ | 1 | $s_{42}$ | $s_{43}$ | $s_{44}$ | $s_{45}$ |
| $x_5$ | 1 | $s_{52}$ | $s_{53}$ | $s_{54}$ | $s_{55}$ |

Wielkość tę można wyrazić również odwołując się do warunku 3.17, tj. lokalnych stosunków szans. Zaczynając od pseudo–przekątnej przylegającej do głównej przekątnej, liczba komórek, które definiują lokalne stosunki szans wynosi $r-2$, a na kolejnych pseudo przekątnych ich liczba wynosi odpowiednio: $r-3$, $r-4$, ..., 1. Zgodnie z omawianą hipotezą stosunki te są symetryczne, tak więc ogółem liczba warunków wynosi:

$$df = 1 + 2 + \ldots + (r-2) = (r-1)(r-2)/2.$$

Dane z tabeli 3.11 pokazują, że dla tablicy ruchliwości zawodowej model ten jest akceptowalny na poziomie istotności 0,01, dla tablicy ruchliwości edukacyjnej byłby akceptowalny nawet na poziomie 0,05, natomiast dla danych panelowych model ten trudno byłoby odrzucić nawet przy bardzo wysokim poziomie istotności. Indeksy rozbieżności wynoszą odpowiednio: 2,4%, 1,0%, i 0,6% co w ostatnim przypadku wskazuje na wręcz idealne dopasowanie do danych.

Tabela 3.21 przedstawia rozkład oczekiwany zgodny z modelem QS dla danych panelowych. W modelu tym — podobnie jak w modelu quasi–niezależności – nie zakłada się nic na temat komórek leżących na głównej przekątnej. Liczebności te odzwierciedlają liczebności z próby. Podobnie, zgodne z danymi empirycznymi są rozkłady brzegowe obydwu zmiennych. Tak jak w modelu symetrii suma liczebności (prawdopodobieństw) oczekiwanych dla każdej pary komórek lokujących się symetrycznie po obydwu stronach głównej przekątnej jest równa analogicznej sumie liczebności (częstości) wyznaczonych dla danych empirycznych. Podsumowując, aby wyznaczyć rozkład oczekiwany metodą największej wiarygodności, muszą być spełnione warunki:

$$\widehat{\pi}_i^X = p_i^X,\ \widehat{\pi}_j^Y = p_j^Y, \tag{3.19}$$

$$\widehat{\pi}_{ii}^{XY} = p_{ii}^{XY}, \tag{3.20}$$

$$\widehat{\pi}_{ij}^{XY} + \widehat{\pi}_{ji}^{XY} = p_{ij}^{XY} + p_{ji}^{XY}. \tag{3.21}$$

Tabela 3.21: Rozkład oczekiwany zgodny z modelem quasi–symetrii dla danych z tabeli 3.6

| Czy przy aktualnym dochodzie netto P Pana(i) gospodarstwo domowe wiąże koniec z końcem? | | | | | | |
|---|---|---|---|---|---|---|
| Odpowiedzi w 2000 roku ($X$) | Odpowiedzi w 2005 roku ($Y$) | | | | | |
| | 1 | 2 | 3 | 4 | 5 | Suma |
| 1. Z wielką trudnością | 271,7 | 168,4 | 115,4 | 23,4 | 3,3 | 582,2 |
| 2. Z trudnością | 102,8 | 134,5 | 169,0 | 43,9 | 6,3 | 456,5 |
| 3. Z pewną trudnością | 55,1 | 132,1 | 248,5 | 98,0 | 12,1 | 545,9 |
| 4. Raczej łatwo | 9,3 | 28,4 | 81,1 | 73,9 | 21,4 | 214,0 |
| 5. Łatwo | 0,9 | 2,6 | 6,5 | 13,8 | 13,5 | 37,2 |
| Suma | 439,7 | 466,0 | 620,4 | 253,1 | 56,6 | 1835,8 |

Tabela 3.22: Parametry modelu quasi–symetrii dla danych z tabeli 3.6

| Czy przy aktualnym dochodzie netto P Pana(i) gospodarstwo domowe wiąże koniec z końcem? | | | | | |
|---|---|---|---|---|---|
| Odpowiedzi w 2000 roku ($X$) | Odpowiedzi w 2005 roku ($Y$) | | | | |
| | 1 | 2 | 3 | 4 | 5 |
| 1. Z wielką trudnością | 6,38 | 2,20 | 0,93 | 0,36 | 0,21 |
| 2. Z trudnością | 2,20 | 1,60 | 1,24 | 0,61 | 0,37 |
| 3. Z pewną trudnością | 0,93 | 1,24 | 1,43 | 1,07 | 0,57 |
| 4. Raczej łatwo | 0,36 | 0,61 | 1,07 | 1,86 | 2,30 |
| 5. Łatwo | 0,21 | 0,37 | 0,57 | 2,30 | 9,56 |

W tabeli 3.22 przedstawione są parametry interakcji dla modelu quasi–symetrii dla danych panelowych, przy czym wykorzystana została parametryzacja odchyleń multiplikatywnych, tj. iloczyn parametrów w każdym wierszu i w każdej kolumnie jest równy 1. Jak łatwo zauważyć parametry te są symetryczne. Parametry na głównej przekątnej odzwierciedlają — podobnie jak w modelu quasi–niezależności — tendencję do wskazywania tej samej odpowiedzi w obydwu punktach czasowych. Zauważmy, że są one większe od 1 dla każdej wartości zmiennej opisującej sytuację materialną gospodarstwa domowego, natomiast najwyższe wartości przyjmują dla kategorii skrajnych.

Łatwo zauważyć, że parametry są generalnie mniejsze dla pseudo–przekątnych oddalonych od głównej przekątnej. Na przykład udzielenie odpowiedzi *z wielką trudnością* w 2000 roku i *łatwo* w 2005 jest prawie pięciokrotnie razy mniej prawdopodobne (1/0,21) aniżeli w hipotetycznej sytuacji braku związku między zmiennymi i – analogicznie – prawie pięciokrotnie razy mniej prawdopodobne jest aby sytuacja gospodarstwa, które w 2000 deklarowało, że wiązało koniec z końcem *łatwo*, pogorszyła się tak znacząco, że w 2005 roku udzielona została odpowiedź *z wielką trudnością*. Na podstawie parametrów interakcji można odtworzyć dowolny stosunek szans. Na przykład:

$$\Theta_{13}^{XY} = \Theta_{31}^{XY} = \frac{0,93 \cdot 0,61}{0,36 \cdot 1,24} = 1,28.$$

Porównując modele symetrii i quasi–symetrii można zauważyć, że pierwszy z nich jest zagnieżdżony w drugim i różnią się one założeniem 3.15, które dotyczy identyczności rozkładów brzegowych obydwu zmiennych. Test warunkowy porównujący modele S i QS jest często stosowany do weryfikacji tej właśnie hipotezy, zwanej w literaturze *homogenicznością brzegową* (*marginal homogenity*) (Caussinus 1965, Haberman 1979). W odniesieniu do tablicy ruchliwości dotyczy ona odpowiedzi na pytania: *czy wykształcenie w pokoleniu ojców ma taką samą strukturę jak w pokoleniu synów*? Natomiast w odniesieniu do danych panelowych: *czy rozkład zmiennej jest identyczny w obydwu badaniach*? W odniesieniu do analizowanych przez nas danych z tabel 3.2, 3.5, 3.6 okazuje się, że założenie to powinno być odrzucone w każdym przypadku na standardowo przyjmowanych poziomach istotności. Dla tablicy ruchliwości zawodowej $G^2 = 362,6 - 18,0 = 344,6$, $df = 5$, dla ruchliwości edukacyjnej, $G^2 = 985,4 - 8,1 = 950,3$, $df = 3$, dla danych panelowych $G^2 = 61,6 - 1,4 = 60,2$, $df = 4$ (w każdym przypadku $p < 0,0001$). Jak widać, założenia o identyczności rozkładów brzegowych nie da się utrzymać. Pomiędzy respondentami a ich ojcami istnieją istotne statystycznie różnice w strukturze przynależności społeczno–zawodowej i strukturze wykształcenia, podobnie — choć tu redukcja jest stosunkowo najmniejsza — należy odrzucić hipotezę, że rozkład zmiennej opisującej sytuację gospodarstw domowych jest identyczny w 2000 i w 2005 roku.

Założenie dotyczące homogeniczności brzegowej można testować również porównując przedstawione wcześniej modele N oraz model NS. W praktyce częściej stosuje się do tych celów model symetrii i quasi-symetrii, co wynika z tego, że modele te są bardziej realistyczne. Jak pamiętamy stosowanie testów warunkowych jest uzasadnione o ile bardziej złożony z porównywanych modeli jest prawdziwy, w przeciwnym razie statystyka $G^2$ może znacząco odbiegać od rozkładu teoretycznego $\chi^2$. Dlatego wykorzystywanie w tym celu modeli N i NS może prowadzić do błędnych konkluzji,

choć zależy to oczywiście od konkretnego przypadku, nie należy traktować tego jako regułę.

Zauważmy, że modele N, NS, QN są zagnieżdżone w modelu quasi–symetrii. Model ten jest dość ogólny, jak się okaże wiele modeli prezentowanych w dalszej części tego rozdziału będzie jego szczególnym przypadkiem, przy czym do opisu związku wykorzystana zostanie informacja o uporządkowaniu kategorii zmiennych. Modele prezentowane do tej pory mogły być stosowane do zmiennych nominalnych.

### 3.2.3 Modele jednakowej interakcji i wierszowo–kolumnowe

W tej części zaprezentowane zostaną modele jednakowej interakcji i modele wierszowo–kolumnowe. Zostały one już zaprezentowane w rozdziale 2, tym razem zostaną omówione w kontekście analizy tablic ruchliwości i danych panelowych. Jak zobaczymy, możliwa jest modyfikacja tych modeli, przy wykorzystaniu informacji o identyczności kategorii zmiennej wierszowej i kolumnowej. Udzielona zostanie również odpowiedź na pytanie, przy jakich założeniach modele te mogą być interpretowane jako szczególne przypadki quasi–symetrii.

Warto przypomnieć, że modele jednakowej interakcji i wierszowo–kolumnowe (podobnie jak modele przedstawione w dalszej części tego rozdziału) na ogół wykorzystują informację o uporządkowaniu kategorii zmiennej wierszowej i kolumnowej. Wykorzystanie informacji o porządkowym charakterze zmiennych, pozwala na formułowanie dodatkowych hipotez. Z drugiej strony należy pamiętać, o ograniczeniach w możliwości stosowania tych modeli. Przykładowo, kategorii społeczno–zawodowych w tablicy ruchliwości (tabela 3.2), nie można uporządkować w niearbitralny sposób, trudno na przykład usytuować „rolników" względem innych kategorii. Dlatego w dalszych ilustracjach empirycznych — poza jednym wyjątkiem[23] — pominiemy dane dotyczące ruchliwości zawodowej.

#### Model jednakowej interakcji

Przypomnijmy, że model jednakowej interakcji (UA), głosi, że wszystkie lokalne stosunki szans, tj. wyróżnione dla sąsiednich kategorii zmiennych porządkowych są identyczne, tj.

$$\Theta_{ij}^{XY} = \delta.$$

Należy zauważyć, że jeśli analizowany jest rozkład dwóch zmiennych o identycznych kategoriach, to hipoteza jednakowej interakcji jest szczególnym typem hipotezy o

[23] Logarytmiczno–multiplikatywny model wierszowo–kolumnowy może być wykorzystywany również do analizy zmiennych nominalnych, jak było sygnalizowane w rozdziale drugim.

quasi–symetrii. Powyższy warunek o identyczności lokalnych stosunków szans implikuje warunek 3.17 definiujący quasi–symetrię. Ponieważ, każdy lokalny stosunek szans wynosi tyle samo w związku z tym równe sobie są również stosunki szans ulokowane symetrycznie względem przekątnej. Można to również przełożyć na parametry modelu logarytmiczno–liniowego. Np. jeśli przyjmiemy identyczne kategorie odniesienia zmiennej wierszowej i kolumnowej tj. $x_a$, $y_a$ to w stosunku do modelu quasi–symetrii 3.18 zachodzi $s_{ij}^{XY} = \delta^{(i-a)(j-a)}$. Ze względów oczywistych zdefiniowany w ten sposób parametr interakcji jest symetryczny, gdyż:

$$\delta^{(i-a)(j-a)} = \delta^{(j-a)(i-a)}$$

Ilustrację dla tego modelu stanowi tabela 3.23. Jak widać, przy przyjęciu takich samych kategorii odniesienia $x_1$, $y_1$ obydwu zmiennych parametry interakcji są symetryczne względem przekątnej. Zmieniłoby się to, gdyby przyjęte zostały różne kategorie odniesienia obydwu zmiennych np. $x_1$, $y_3$, niemniej model taki pozostawałby modelem quasi–symetrycznym, gdyż symetryczne byłyby stosunki szans, tj. spełniony byłby warunek 3.17.

Tabela 3.23: Ilustracja modelu *jednakowej interakcji* — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\delta$ | $\delta^2$ | $\delta^3$ | $\delta^4$ |
| $x_3$ | 1 | $\delta^2$ | $\delta^4$ | $\delta^6$ | $\delta^8$ |
| $x_4$ | 1 | $\delta^3$ | $\delta^6$ | $\delta^9$ | $\delta^{12}$ |
| $x_5$ | 1 | $\delta^4$ | $\delta^8$ | $\delta^{12}$ | $\delta^{16}$ |

Przypomnijmy, że model ten posiada tylko jeden parametr więcej w stosunku do modelu niezależności, jego liczba stopni swobody wynosi $df = (r-1)^2-1$. Taka liczba stopni swobody może być również uzasadniona równością lokalnych stosunków szans, których dla tabeli o wymiarach $r$ na $r$ można wyznaczyć $(r-1)^2$.

W tabeli 3.24 przedstawione zostały wyniki dopasowania tego modelu do danych dla tablic 3.5 oraz 3.6. W przypadku tablicy ruchliwości edukacyjnej model jednakowej interakcji powinien być odrzucony na poziomie istotności 0,01. Jeśli chodzi o dane panelowe model ten może być zaakceptowany na poziomie istotności 0,05. Parametr interakcji — czyli lokalny stosunek szans — wynosi w przypadku danych panelowych $\delta = 1,7$. Oznacza to na przykład, że proporcja liczby gospodarstw, które w 2005 roku radziły sobie *z trudnością*, do liczby gospodarstw, które koniec z końcem wiązały

Tabela 3.24: Wyniki weryfikacji modeli jednakowej interakcji i wierszowo–kolumnowych dla tabel 3.5 i 3.6

| Model | df | $\chi^2$ | $G^2$ | $\Delta$ |
|---|---|---|---|---|
| Wyniki dla tabeli (3.5) (tablica ruchliwości edukacyjnej) | | | | |
| UA | 8 | 22,4 ($p = 0,0042$) | 24,9 ($p = 0,0016$) | 4,1 |
| QUA | 4 | 10,5 ($p = 0,0332$) | 11,8 ($p = 0,0192$) | 1,8 |
| R | 6 | 21,9 ($p = 0,0013$) | 24,0 ($p = 0,0005$) | 4,1 |
| QR | 2 | 1,1 ($p = 0,5862$) | 1,6 ($p = 0,4511$) | 0,3 |
| C | 6 | 10,6 ($p = 0,0999$) | 11,6 ($p = 0,0718$) | 2,2 |
| QC | 2 | 6,9 ($p = 0,0310$) | 7,1 ($p = 0,0287$) | 1,5 |
| RC1 | 4 | 6,1 ($p = 0,1906$) | 6,3 ($p = 0,1790$) | 1,6 |
| (R=C)1 | 6 | 20,6 ($p = 0,0021$) | 23,1 ($p = 0,0008$) | 3,6 |
| RC2 | 4 | 8,1 ($p = 0,0878$) | 6,8 ($p = 0,1447$) | 1,7 |
| (R=C)2 | 6 | 20,1 ($p = 0,0027$) | 22,6 ($p = 0,0010$) | 3,4 |
| Wyniki dla tabeli (3.6) (dane panelowe) | | | | |
| UA | 15 | 24,5 ($p = 0,0569$) | 23,6 ($p = 0,0729$) | 3,9 |
| QUA | 10 | 9,6 ($p = 0,4729$) | 9,4 ($p = 0,4928$) | 1,4 |
| R | 12 | 20,0 ($p = 0,0673$) | 15,9 ($p = 0,1947$) | 2,8 |
| QR | 7 | 6,2 ($p = 0,5192$) | 5,6 ($p = 0,5886$) | 1,2 |
| C | 12 | 23,2 ($p = 0,0259$) | 20,7 ($p = 0,0554$) | 3,3 |
| QC | 7 | 8,7 ($p = 0,2716$) | 8,6 ($p = 0,2834$) | 1,3 |
| RC1 | 9 | 19,7 ($p = 0,0199$) | 13,9 ($p = 0,1278$) | 2,2 |
| QRC1 | 4 | 0,8 ($p = 0,9390$) | 0,8 ($p = 0,9412$) | 0,3 |
| (R=C)1 | 12 | 20,4 ($p = 0,0591$) | 16,5 ($p = 0,1709$) | 2,5 |
| Q(R=C)1 | 7 | 1,9 ($p = 0,9654$) | 1,9 ($p = 0,9659$) | 0,7 |
| RC2 | 9 | 19,4 ($p = 0,0223$) | 13,8 ($p = 0,1304$) | 2,2 |
| QRC2 | 4 | 0,6 ($p = 0,9574$) | 0,7 ($p = 0,9566$) | 0,4 |
| (R=C)2 | 12 | 20,5 ($p = 0,0586$) | 16,1 ($p = 0,1849$) | 2,5 |
| Q(R=C)2 | 7 | 1,5 ($p = 0,9836$) | 1,4 ($p = 0,9841$) | 0,7 |

*z wielką z trudnością* była 1,7 razy większa wśród gospodarstw, które w 2000 roku radziły sobie *z trudnością* niż wśród tych radziły sobie *z wielką trudnością*.

Łatwo można wyznaczyć również stosunek szans dla kategorii zmiennych porządkowych, które nie sąsiadują ze sobą. Przykładowo, porównajmy te same kategorie gospodarstw z 2000 roku co poprzednio, tym razem ze względu na to jak często w

2005 roku wskazują dwie skrajne kategorie. Ponieważ odpowiedzi te dzielą cztery kategorie, lokalny stosunek szans należy podnieść do potęgi czwartej, tj. porównując proporcję liczby gospodarstw, które radziły sobie *łatwo* do liczby gospodarstw *z wielką z trudnością* byłaby ona 8,34 (tj. $1,7^4$) razy większa wśród gospodarstw, które w 2000 roku radziły sobie *z trudnością* w porównaniu do gospodarstw, które radziły sobie *z wielką trudnością*.

Model jednakowej interakcji — jeśli odnosi się go do danych gdzie kategorie zmiennej wierszowej i kolumnowej są identyczne — modyfikowany jest często przez uwzględnienie specyfiki komórek na głównej przekątnej. Podobnie jak w przypadku modelu niezależności stochastycznej i modelu QN. Warto przypomnieć, że model taki interpretowany był w kategoriach wyróżnienia dodatkowej podzbiorowości: pierwszej ($Z = 1$), do której należą wyłącznie osoby niemobilne i dwóch, w których związek jest opisywany w kategoriach niezależności, przy czym różniły się one ze względu na to czy wyklucza się możliwość dziedziczenia pozycji ($Z = 2$), czy też nie ($Z = 3$).

Taki model można uogólnić, jeśli nie będziemy przesądzać czy mobilność jest opisywana za pomocą niezależności. Można przyjąć, że jest to proces opisywany inaczej, w tym przypadku za pomocą hipotezy o jednakowej interakcji. Hipoteza taka jest wówczas szczególnym przypadkiem przywoływanego wcześniej modelu $MS$, gdzie dzieli zbiorowość na osoby niemobilne ($Z = 1$) i mobilne ($Z = 2$ oraz $Z = 3$).

Model jednakowej interakcji uwzględniający specyfikę komórek położonych na głównej przekątnej będzie oznaczany jako QUA. Trzy wymienione powyżej zbiorowości zostały zilustrowane za pomocą tabeli 3.25. Odsetki poszczególnych podzbiorowości można oznaczyć, tak jak w przypadku quasi–niezależności:$\beta_1$, $\beta_2$ oraz $\beta_3$, analogicznie odpowiednie prawdopodobieństwo rozkładu łącznego zgodne z modelem QUA dane jest formułą 3.7. Warto zauważyć, że nie musi zachodzić $a_i^X = b_i^X$, ani też $a_j^Y = b_j^Y$, natomiast parametr interakcji opisujący lokalny stosunek szans ($\delta$), musi być taki sam dla podzbiorowości $Z = 2$ oraz $Z = 3$. Dla podzbiorowości $Z = 2$ wszystkie lokalne stosunki szans są równe $\delta$, a nie–lokalne stosunki szans są opisywane przez wartość $\delta$ podniesioną do odpowiedniej potęgi, która wynika z tego, ile kategorii dzieli porównywane wartości zmiennej porządkowej. Podobnie dzieje się dla podzbiorowości $Z = 3$, przy czym nie dotyczy to stosunków szans, które obejmują komórki na głównej przekątnej.

Inną — parametryczną — ilustracją tego samego modelu jest tablica 3.26, gdzie poza parametrami $\delta$ opisującymi jednakową interakcję, znajdują się parametry $q_i$ opisujące kolejne komórki głównej przekątnej. W tabeli 3.27 przedstawiony został rozkład oczekiwany dla danych panelowych zgodny z tym modelem. Model QUA wymaga $r$ niezależnych parametrów więcej niż model UA, tj. dla każdej komórki

Tabela 3.25: Ilustracja modelu QUA z uwzględnieniem podziału zbiorowości ze względu na zmienną $Z$

| $Z=1$ | | | | | |
|---|---|---|---|---|---|
| $X\backslash Y$ | 1 | 2 | 3 | 4 | 5 |
| 1 | $w_1$ | 0 | 0 | 0 | 0 |
| 2 | 0 | $w_2$ | 0 | 0 | 0 |
| 3 | 0 | 0 | $w_3$ | 0 | 0 |
| 4 | 0 | 0 | 0 | $w_4$ | 0 |
| 5 | 0 | 0 | 0 | 0 | $w_5$ |
| $Z=2$ | | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | 5 |
| 1 | $a_1^X a_1^Y$ | $a_1^X a_2^Y$ | $a_1^X a_3^Y$ | $a_1^X a_4^Y$ | $a_1^X a_5^Y$ |
| 2 | $a_2^X a_1^Y$ | $a_2^X a_2^Y \delta$ | $a_2^X a_3^Y \delta^2$ | $a_2^X a_4^Y \delta^3$ | $a_2^X a_5^Y \delta^4$ |
| 3 | $a_3^X a_1^Y$ | $a_3^X a_2^Y \delta^2$ | $a_3^X a_3^Y \delta^4$ | $a_3^X a_4^Y \delta^6$ | $a_3^X a_5^Y \delta^8$ |
| 4 | $a_4^X a_1^Y$ | $a_4^X a_2^Y \delta^3$ | $a_4^X a_3^Y \delta^6$ | $a_4^X a_4^Y \delta^9$ | $a_4^X a_5^Y \delta^{12}$ |
| 5 | $a_5^X a_1^Y$ | $a_5^X a_2^Y \delta^4$ | $a_5^X a_3^Y \delta^8$ | $a_5^X a_4^Y \delta^{12}$ | $a_5^X a_5^Y \delta^{16}$ |
| $Z=3$ | | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | $b_1^X b_2^Y$ | $b_1^X b_3^Y$ | $b_1^X b_4^Y$ | $b_1^X b_5^Y$ |
| 2 | $b_2^X b_1^Y$ | 0 | $b_2^X b_3^Y \delta^2$ | $b_2^X b_4^Y \delta^3$ | $b_2^X b_5^Y \delta^4$ |
| 3 | $b_3^X b_1^Y$ | $b_3^X b_2^Y \delta^2$ | 0 | $b_3^X b_4^Y \delta^6$ | $b_3^X b_5^Y \delta^8$ |
| 4 | $b_4^X b_1^Y$ | $b_4^X b_2^Y \delta^3$ | $b_4^X b_3^Y \delta^6$ | 0 | $b_4^X b_5^Y \delta^{12}$ |
| 5 | $b_5^X b_1^Y$ | $b_5^X b_2^Y \delta^4$ | $b_5^X b_3^Y \delta^8$ | $b_5^X b_4^Y \delta^{12}$ | 0 |

Tabela 3.26: Ilustracja modelu *jednakowej interakcji* przy uwzględnieniu specyfiki komórek na głównej przekątnej — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $q_1$ | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\delta q_2$ | $\delta^2$ | $\delta^3$ | $\delta^4$ |
| $x_3$ | 1 | $\delta^2$ | $\delta^4 q_3$ | $\delta^6$ | $\delta^8$ |
| $x_4$ | 1 | $\delta^3$ | $\delta^6$ | $\delta^9 q_4$ | $\delta^{12}$ |
| $x_5$ | 1 | $\delta^4$ | $\delta^8$ | $\delta^{12}$ | $\delta^{16} q_5$ |

opisującej główną przekątną konieczne jest uwzględnienie oddzielnego parametru $q_i$ — tak jak ilustruje to tabela 3.26. Jego liczba stopni swobody wynosi więc:

$$df = (r-1)^2 - 1 - r = r(r-3)$$

Hipotezę QUA można sformułować odwołując się wyłącznie do lokalnych stosunków szans. Formuła taka zamieszczona została w Aneksie, tam również można znaleźć bardziej precyzyjne wyjaśnienie liczby stosunków szans dla tego modelu.

Tabela 3.27: Rozkład oczekiwany zgodny z QUA dla danych z tabeli 3.6

| Czy przy aktualnym dochodzie netto Pana(i) gospodarstwo domowe wiąże koniec z końcem? | | | | | | |
|---|---|---|---|---|---|---|
| Odpowiedzi w 2000 roku ($X$) | Odpowiedzi w 2005 roku ($Y$) | | | | | Suma |
| | 1 | 2 | 3 | 4 | 5 | |
| 1. Z wielką trudnością | 271,7 | 170,8 | 116,1 | 21,9 | 1,7 | 582,2 |
| 2. Z trudnością | 104,1 | 134,5 | 163,7 | 48,4 | 5,8 | 456,5 |
| 3. Z pewną trudnością | 54,7 | 126,7 | 248,5 | 97,7 | 18,2 | 545,8 |
| 4. Raczej łatwo | 8,7 | 31,6 | 82,4 | 73,9 | 17,4 | 214,0 |
| 5. Łatwo | 0,4 | 2,4 | 9,8 | 11,1 | 13,5 | 37,2 |
| Suma | 439,6 | 466,0 | 620,5 | 253,0 | 56,6 | 1835,7 |

Oszacowanie tego modelu metodą największej wiarygodności wymaga, aby spełnione były warunki 2.13-2.15 formułowane dla zwykłego modelu jednakowej interakcji, z drugiej strony — podobnie jak w modelu quasi–niezależności — liczebności oczekiwane komórek na głównej przekątnej muszą odzwierciedlać liczebności empiryczne.

Wyniki weryfikacji przedstawione w tabeli 3.24 wskazują, że w odniesieniu do tablicy ruchliwości edukacyjnej model jest akceptowalny na poziomie istotności równy 0,01. Parametr opisujący jednakową interakcję wynosi $\delta = 1,60$, natomiast parametry głównej przekątnej są równe odpowiednio $q_1 = 1,91$, $q_2 = 1,11$, $q_3 = 1,02$, $q_4 = 1,58$. Wielkości tych parametrów są nieco inne, niż w przypadku modelu quasi–niezależności. Co istotniejsze, zmienia się również interpretacja tych parametrów. Wskazują one, na ile dziedzicznie danej pozycji jest bardziej prawdopodobne aniżeli w hipotetycznej sytuacji modelu jednakowej interakcji, nie zaś niezależności.

Statystyczne uzasadnienie uwzględnienia tych parametrów jest kontrowersyjne. Test warunkowy porównujący modele UA i QUA daje wartość statystyki $G^2 = 24,9 - 11,8 = 13,1$, $df = 4$, $p = 0,0106$. Redukcja ta nie jest tak duża jak w przypadku modelu quasi-niezależności. Wynika to z tego, że jeśli w modelu jednakowej

interakcji $\delta > 1$, to wszystkie lokalne stosunki szans są większe od 1, a w konsekwencji wyższe wartości jednej zmiennej relatywnie częściej występują z wyższymi wartościami drugiej zmiennej a niższe z niższymi. W konsekwencji zgodnie z tym modelem, komórki na głównej przekątnej mają relatywnie większe częstości i modyfikacja modelu przez wprowadzenie parametrów głównej przekątnej nie poprawia tak znacząco dopasowania jak w przypadku modelu niezależności.

Jeśli chodzi o dane panelowe dopasowanie tego modelu do danych jest bardzo dobre, choć również w tym przypadku uwzględnienie parametrów głównej przekątnej jest dyskusyjne: $G^2 = 23,6 - 9,4 = 14,2$, $df = 5$, $p = 0,0144$. Parametr jednakowej interakcji wynosi: $\delta = 1,55$, natomiast parametry głównej przekątnej są równe: $q_1 = 1,50$, $q_2 = 0,87$, $q_3 = 1,17$, $q_4 = 1,23$, $q_4 = 2,65$.

### Model wierszowo–kolumnowy I — wersja symetryczna

Jak zostało powiedziane w rozdziale 2 hipoteza dotycząca modelu wierszowo–kolumnowego RC głosi, że lokalny stosunek szans — dotyczący zmiennych porządkowych — zależy z jednej strony od wartości zmiennej $X$, z drugiej strony od wartości zmiennej $Y$, a dokładniej:

$$\Theta_{ij}^{XY} = \delta_{i.}\delta_{.j}$$

Wyniki z tabeli 3.24 pokazują, że model ten jest akceptowalny zarówno dla tablicy ruchliwości edukacyjnej jak i danych panelowych. Zauważmy, że w przypadku, gdy mamy do czynienia z tablicą o takich samych kategoriach zmiennej wierszowej i kolumnowej, model ten można uprościć, przyjmując dodatkowo, że:

$$\delta_{i.} = \delta_{.i} \tag{3.22}$$

Tabela 3.28: Ilustracja symetrycznego modelu wierszowo–kolumnowego I typu — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\psi_2^2$ | $\psi_2^2 \cdot \psi_3$ | $\psi_2^3 \cdot \psi_4$ | $\psi_2^4 \cdot \psi_5$ |
| $x_3$ | 1 | $\psi_3 \cdot \psi_2^2$ | $\psi_3^4$ | $\psi_3^3 \cdot \psi_4^2$ | $\psi_3^4 \cdot \psi_5^2$ |
| $x_4$ | 1 | $\psi_4 \cdot \psi_2^3$ | $\psi_4^2 \cdot \psi_3^3$ | $\psi_4^6$ | $\psi_4^4 \cdot \psi_5^3$ |
| $x_5$ | 1 | $\psi_5 \cdot \psi_2^4$ | $\psi_5^2 \cdot \psi_3^4$ | $\psi_5^3 \cdot \psi_4^4$ | $\psi_5^8$ |

Hipotezę taką będziemy oznaczać $(R = C)1$ i może być ona postrzegana jako szczególny przypadek modelu quasi–symetrii. Parametry interakcji są symetryczne,

jeśli kategoriami odniesienia są odpowiadające sobie kategorie $x_a$ oraz $y_a$ obydwu zmiennych:

$$\pi_{ij}^{XY} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \psi_i^{(j-a)} \cdot \psi_j^{(i-a)}, \tag{3.23}$$

przy czym można założyć, że $\psi_1$=1. Tabela 3.28 ilustruje symetryczny model wierszowo–kolumnowy tego typu. Przypomnijmy, że w zwykłym modelu wierszowo–kolumnowym spełniony był warunek 2.31:

$$\frac{\Theta_{aj}^{XY}}{\Theta_{bj}^{XY}} = \frac{\delta_{a\cdot}}{\delta_{b\cdot}},$$

czyli iloraz dwóch lokalnych stosunków szans wyróżnionych dla różnych kategorii zmiennej wierszowej $x_a$, $x_b$ o ile dotyczył tej samej kolumny $y_j$. Ponieważ w symetrycznym modelu wierszowo–kolumnowym odpowiednie stosunki szans są symetryczne, tj. $\Theta_{aj}^{XY} = \Theta_{bj}^{XY}$, dodatkowo zachodzą następujące równości:

$$\frac{\Theta_{aj}^{XY}}{\Theta_{bj}^{XY}} = \frac{\Theta_{ia}^{XY}}{\Theta_{ib}^{XY}} = \frac{\Theta_{ka}^{XY}}{\Theta_{bk}^{XY}} = \frac{\Theta_{al}^{XY}}{\Theta_{lb}^{XY}} = \frac{\delta_{a\cdot}}{\delta_{b\cdot}} \tag{3.24}$$

Okazuje się, że równe sobie są wszystkie ilorazy stosunków szans, o ile w liczniku stosunek szans dotyczy kategorii $a$, natomiast stosunek szans w mianowniku dotyczy kategorii $b$, zaś pozostałe dwa indeksy dotyczą tej samej kategorii obojętne czy zmiennej wierszowej czy też zmiennej kolumnowej.

Aby wyznaczyć liczbę stopni swobody dla tego modelu, warto przeanalizować tabelę 3.29, w której przedstawione zostały stosunki szans wyznaczone na podstawie tabeli 3.28. Na przecięciu $i$-tego wiersza oraz $j$-tej kolumny umieszczony został lokalny stosunek szans $\Theta_{ij}^{XY}$. W stosunku do tabeli 3.28 $\delta_1 = \psi_{i+1}/\psi_i$, np. $\delta_1 = \psi_2/\psi_1$, $\delta_2 = \psi_3/\psi_2$, itd. Jak widać, stosunki szans są symetryczne, co przekłada się na 6 równości, a bardziej ogólnie liczba warunków tego typu wynosi $(r-2)(r-1)/2$, tak jak w przypadku quasi–symetrii. Ponadto zachodzą warunki, określone przez model wierszowo–kolumnowy:

$$\frac{\Theta_{11}^{XY}}{\Theta_{21}^{XY}} = \frac{\Theta_{12}^{XY}}{\Theta_{22}^{XY}} = \frac{\Theta_{13}^{XY}}{\Theta_{23}^{XY}} = \frac{\Theta_{14}^{XY}}{\Theta_{24}^{XY}}, \quad \frac{\Theta_{22}^{XY}}{\Theta_{32}^{XY}} = \frac{\Theta_{23}^{XY}}{\Theta_{33}^{XY}} = \frac{\Theta_{24}^{XY}}{\Theta_{34}^{XY}}, \quad \frac{\Theta_{33}^{XY}}{\Theta_{43}^{XY}} = \frac{\Theta_{34}^{XY}}{\Theta_{44}^{XY}},$$

Co daje sześć ograniczeń, a jeśli mamy $r$ kategorii, ich liczba wynosi:

$$(r-2) + (r-3) \ldots + 1 = \frac{(r-2)(r-1)}{2}.$$

Nie jest na przykład konieczne zakładanie dodatkowo, że:

$$\frac{\Theta_{31}^{XY}}{\Theta_{41}^{XY}} = \frac{\Theta_{32}^{XY}}{\Theta_{42}^{XY}},$$

gdyż wynika to z symetrii i przytoczonego powyżej warunku:

$$\frac{\Theta_{13}^{XY}}{\Theta_{23}^{XY}} = \frac{\Theta_{14}^{XY}}{\Theta_{24}^{XY}}.$$

Liczba stopni swobody dla symetrycznej postaci modelu wierszowo–kolumnowego wynosi więc:

$$df = \frac{(r-2)(r-1)}{2} + \frac{(r-2)(r-1)}{2} = (r-2)(r-1).$$

Tabela 3.29: Lokalne stosunki szans wyznaczone dla tabeli 3.28[a]

| $X \backslash Y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | $\delta_1^2$ | $\delta_1\delta_2$ | $\delta_1\delta_3$ | $\delta_1\delta_4$ |
| 2 | $\delta_2\delta_1$ | $\delta_2^2$ | $\delta_2\delta_3$ | $\delta_2\delta_4$ |
| 3 | $\delta_3\delta_1$ | $\delta_3\delta_2$ | $\delta_3^2$ | $\delta_3\delta_4$ |
| 4 | $\delta_4\delta_1$ | $\delta_4\delta_2$ | $\delta_4\delta_3$ | $\delta_4^2$ |

[a]W stosunku do tabeli 3.28 $\delta_1 = \psi_{i+1}/\psi_i$, np. $\delta_2 = \psi_3/\psi_2$.

Aby wyznaczyć rozkład oczekiwany metodą największej wiarygodności dla symetrycznego modelu wierszowo-kolumnowego rozkłady brzegowe szacujemy na podstawie danych empirycznych. Ponadto należy sformułować warunek:

$$\sum_{i=1}^{r} i \cdot \widehat{\pi}_{ia}^{XY} + \sum_{j=1}^{r} j \cdot \widehat{\pi}_{aj}^{XY} = \sum_{i=1}^{r} i \cdot p_{ia}^{XY} + \sum_{j=1}^{r} j \cdot p_{aj}^{XY}, \tag{3.25}$$

dla każdej kategorii $a$ zmiennej $X$ oraz zmiennej $Y$. Jak widać, warunków tych jest mniej w porównaniu do 2.37–2.38, gdyż wersja symetryczna jest hipotezą prostszą niż jej asymetryczny odpowiednik, niemniej widoczne są pewne analogie pomiędzy tymi formułami. Rozkład oczekiwany zgodny z tą hipotezą przedstawiony został w tablicy 3.30.

Jak można zauważyć zachodzą warunki przedstawione w formule 3.24, na przykład:

$$\frac{\Theta_{12}^{XY}}{\Theta_{42}^{XY}} = \frac{(169,1 \cdot 167,4)/(125,4 \cdot 133,3)}{(30,5 \cdot 9,5)/(85,5 \cdot 1,9)} = \frac{\Theta_{31}^{XY}}{\Theta_{43}^{XY}} = \frac{(60,2 \cdot 30,5)/(131,2 \cdot 7,4)}{(85,5 \cdot 14,8)/(67,9 \cdot 9,5)}$$

Jak pokazuje tabela 3.24 symetryczny model wierszowo–kolumnowy (R=C)1 jest dobrze dopasowany do danych panelowych, natomiast powinien być odrzucony na poziomie istotności 0,01 w odniesieniu do danych ruchliwości edukacyjnej. Łatwo zauważyć, że wersja symetryczna (R=C)1 jest zagnieżdżona w asymetrycznej wersji modelu RC i różni je założenie o symetrii lokalnych stosunków szans. W przypadku danych panelowych wyniki testu warunkowego $G^2 = 2,61$, $df = 3$ ($p = 0,45$),

Tabela 3.30: Rozkład oczekiwany zgodny z $(R = C)1$ dla danych z tabeli 3.6

| Czy przy aktualnym dochodzie netto Pana(i) gospodarstwo domowe wiąże koniec z końcem? | | | | | | |
|---|---|---|---|---|---|---|
| Odpowiedzi w 2000 roku ($X$) | Odpowiedzi w 2005 roku ($Y$) | | | | | Suma |
| | 1 | 2 | 3 | 4 | 5 | |
| 1. Z wielką trudnością | 268,1 | 169,1 | 125,4 | 18,7 | 0,9 | 582,2 |
| 2. Z trudnością | 103,7 | 133,3 | 167,4 | 47,5 | 4,6 | 456,5 |
| 3. Z pewną trudnością | 60,2 | 131,2 | 232,6 | 104,2 | 17,7 | 545,9 |
| 4. Raczej łatwo | 7,4 | 30,5 | 85,5 | 67,9 | 22,7 | 214,0 |
| 5. Łatwo | 0,2 | 1,9 | 9,5 | 14,8 | 10,8 | 37,2 |
| Suma | 439,6 | 466,0 | 620,4 | 253,1 | 56,7 | 1835,8 |

pokazują, że model symetryczny nie jest dopasowany istotnie gorzej aniżeli wersja asymetryczna, natomiast w odniesieniu do danych ruchliwości dodatkowe założenie o symetrii interakcji musi zostać odrzucone: $G^2 = 16,8$, $df = 2$ $(p = 0,0002)$.

Model wierszowo–kolumnowy w wersji symetrycznej bądź asymetrycznej można zmodyfikować uwzględniając specyfikę głównej przekątnej, będą one oznaczane jako $Q(R = C)1$ oraz $QRC1$. Ilustracje tych modeli zawierają tabele 3.31 oraz 3.32. Modele te można opisać - podobnie jak w przypadku modelu quasi–niezależności i modelu QUA - uwzględniając dodatkową podzbiorowość osób niemobilnych. Dla pozostałych osób, związek opisywany jest za pomocą modelu wierszowo–kolumnowego w symetryczne bądź asymetrycznej wersji, przy czym dla części z tych osób wykluczamy możliwość wystąpienia tej samej wartości obydwu zmiennych $X$ i $Y$. W jednym i w drugim przypadku uwzględnienie parametrów głównej przekątnej wymaga dodatkowo $r$ niezależnych parametrów związanych z komórkami na przekątnej, tj $q_i$. Liczba stopni swobody będzie wynosiła odpowiednio:

$$df = (r - 2)(r - 1) - r,$$

dla wersji symetrycznej, oraz

$$df = (r - 2)^2 - r$$

dla wersji asymetrycznej wersji modelu wierszowo–kolumnowego. Formuły dotyczące obydwu tych modeli i uzasadnienie ich liczby stopni swobody zostało przedstawione w Aneksie.

Należy zauważyć, że modele mają zastosowanie jeśli mamy do czynienia z tabelą o wymiarach 5 x 5 lub większą. Jeśli tabela ma wymiary 4 x 4, wówczas model $QRC1$

Tabela 3.31: Ilustracja symetrycznego modelu wierszowo–kolumnowego I typu z uwzględnieniem specyfiki głównej przekątnej — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $q_1$ | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\psi_2^2 q_2$ | $\psi_2^2 \cdot \psi_3$ | $\psi_2^3 \cdot \psi_4$ | $\psi_2^4 \cdot \psi_5$ |
| $x_3$ | 1 | $\psi_3 \cdot \psi_2^2$ | $\psi_3^4 q_3$ | $\psi_3^3 \cdot \psi_4^2$ | $\psi_3^4 \cdot \psi_5^2$ |
| $x_4$ | 1 | $\psi_4 \cdot \psi_2^3$ | $\psi_4^2 \cdot \psi_3^3$ | $\psi_4^6 q_4$ | $\psi_4^4 \cdot \psi_5^3$ |
| $x_5$ | 1 | $\psi_5 \cdot \psi_2^4$ | $\psi_5^2 \cdot \psi_3^4$ | $\psi_5^3 \cdot \psi_4^4$ | $\psi_5^8 q_5$ |

Tabela 3.32: Rozkład łączny ilustrujący asymetryczny model wierszowo–kolumnowy I typu z uwzględnieniem specyfiki głównej przekątnej

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $q_1$ | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\psi_2 \cdot \phi_2 \cdot q_2$ | $\psi_2^2 \cdot \phi_3$ | $\psi_2^3 \cdot \phi_4$ | $\psi_2^4 \cdot \phi_5$ |
| $x_3$ | 1 | $\psi_3 \cdot \phi_2^2$ | $\psi_3^2 \cdot \phi_3^2 \cdot q_3$ | $\psi_3^3 \cdot \phi_4^2$ | $\psi_3^4 \cdot \phi_5^2$ |
| $x_4$ | 1 | $\psi_4 \cdot \phi_2^3$ | $\psi_4^2 \cdot \phi_3^3$ | $\psi_4^3 \cdot \phi_4^3 \cdot q_4$ | $\psi_4^4 \cdot \phi_5^3$ |
| $x_5$ | 1 | $\psi_5 \cdot \phi_2^4$ | $\psi_5^2 \cdot \phi_3^4$ | $\psi_5^3 \cdot \phi_4^4$ | $\psi_5^4 \cdot \phi_5^4 \cdot q_5$ |

miałby zero stopni swobody. W przypadku modelu $Q(R = C)1$ jego liczba stopni swobody wynosiłaby teoretycznie 1, natomiast model quasi–symetrii dla tablicy o takich samych rozmiarach miałby 3 stopnie swobody. Jest to sprzeczne, jeśli weźmiemy pod uwagę, że model $Q(R = C)1$ jest zagnieżdżony w modelu QS. Dlatego, formułowanie hipotezy tego typu nie ma sensu w odniesieniu do tablicy o takich wymiarach. Model taki pełniłby jedynie funkcję opisową.

Dlatego powyższe modele zostały odniesione jedynie w stosunku danych panelowych. Wyniki weryfikacji z tabeli 3.24 pokazują, że dopasowanie zarówno dla symetrycznej jak też asymetrycznej wersji modelu uwzględniającego specyfikę głównej przekątnej jest znakomite. Na przykład jedynie około 0,7% próby gospodarstw jest zaklasyfikowana niezgodnie z modelem $Q(R = C)1$ i 0,3% – z modelem $QRC1$.

**Niesymetryczne modele wierszowe i kolumnowe.**

Omówione powyżej modele wierszowo-kolumnowe mogą być pomocne w odniesieniu do analizy tablic ruchliwości i danych panelowych, gdy analizowane zmienne mierzone są na skali porządkowej. Jak zostało pokazane, przyjęcie dodatkowych założeń

pozwala na uwzględnienie specyfiki zmiennych o takich samych kategoriach. Z jednej strony możliwe jest przyjęcie dodatkowych założeń dotyczących symetrii modelu, ponadto możliwe jest uwzględnienie specyfiki komórek położonych na głównej przekątnej.

Warto w tym miejscu zaznaczyć, że nie zostały omówione wszystkie modele wierszowo–kolumnowe, jakie mogą być wykorzystane w odniesieniu do tablic ruchliwości i danych panelowych. Możliwe jest sformułowanie w odniesieniu do danych tego typu hipotez o efekcie wierszowym lub kolumnowym omawianych w rozdziale drugim. Nie są to co prawda modele specyficzne dla danych o takich samych kategoriach obydwu zmiennych — podobnie jest w przypadku asymetrycznej wersji modelu wierszowo–kolumnowego — niemniej hipotezy te również wykorzystywane są do analizy danych tego typu (Goodman 1979*b*, Hout 1983, Breen 1985). W tym miejscu modele te nie będą opisane szczegółowo, omówione zostanie dopasowanie wierszowego i kolumnowego w odniesieniu do analizowanych tabel (dopasowanie modeli wierszowo–kolumnowych RC1, QRC1 zostało opisane powyżej).

W tabeli 3.24 przedstawione zostały wyniki weryfikacji tych hipotez. W odniesieniu do tablicy ruchliwości edukacyjnej model wierszowy (R) nie jest akceptowalny. Inaczej jest z modelem kolumnowym, porównanie za pomocą testu warunkowego z modelem jednakowej interakcji pokazuje, że jest on dopasowany istotnie lepiej: $G^2 = 13,3$, $df = 2$ ($p < 0,0013$). Odpowiednie parametry kolumnowe wynoszą: $\phi_1 = 1$, $\phi_2 = 2,9$, $\phi_3 = 4,6$, $\phi_4 = 7,9$. Wartości parametrów są monotoniczne względem kategorii zmiennej wierszowej. Gdybyśmy chcieli zrekonstruować lokalne stosunki szans, tj. porównać sąsiednie kategorie, to okazuje się, że relatywnie największe są lokalne stosunki szans porównujące osoby z wykształceniem podstawowym i niepełnym średnim, tj. kategorie te najbardziej różnią się od siebie ze względu na wykształcenie ojca. W odniesieniu do danych panelowych modele wierszowe i kolumnowe — w porównaniu do modelu jednakowej interakcji — nie przynoszą poprawy istotnej statystycznie na tym poziomie istotności 0,05.

W modelach wierszowych i kolumnowych możliwe jest również uwzględnienie parametrów opisujących główną przekątną. Za każdym razem modyfikacja taka wymaga uwzględnienia $r$ dodatkowych parametrów. Okazuje się, że dla tablicy ruchliwości model wierszowy uwzględniający parametry głównej przekątnej (QR) ma dopasowanie niemal idealne. W odniesieniu do modelu kolumnowego (QC) poprawa dopasowania nie jest istotna statystycznie. Jeśli chodzi o dane panelowe dodanie parametrów głównej przekątnej nie prowadzi do istotnej statystycznie poprawy dopasowania modelu wierszowego na poziomie istotności 0,05. Inaczej jest w przypadku modyfikacji mode-

lu kolumnowego. Wyniki testu warunkowego porównującego modele C i QC wynoszą $G^2 = 12,1$, $df = 5$ $(p = 0,0334)$.

### Symetryczny model wierszowo–kolumnowy II

Tak jak zostało przedstawione w rozdziale drugim, hipoteza związana z logarytmiczno–multiplikatywną wersją modelu wierszowo–kolumnowego głosi, że wielkość lokalnego stosunku szans uwzględnia odległości pomiędzy kategoriami zmiennej, tj.

$$\Theta_{ij}^{XY} = \delta^{(u_{(i+1)}-u_i)(v_{(j+1)}-v_j)} \text{ dla każdej pary } i,\ j, \text{ takich, że } i \leq r-1,\ j \leq r-1. \tag{3.26}$$

Wielkości $u_i$ oraz $v_j$ są parametrami szacowanymi w modelu i dają pewien rodzaj skalowania wartości zmiennych $X$ oraz $Y$. Model ten może być zaakceptowany zarówno dla tablicy ruchliwości edukacyjnej jak i danych panelowych. W pierwszym przypadku skalowanie dla kategorii wykształcenia ojca jest następujące: $u_1 = -0,60$ dla wykształcenia podstawowego, $u_2 = -0,30$ dla wykształcenia niepełnego średniego, $u_3 = 0,19$ dla wykształcenia średniego, $u_4 = 0,71$ wykształcenia wyższego lub niepełnego wyższego, natomiast jeśli chodzi o wykształcenie respondenta odpowiednie parametry są równe $v_1 = -0,79$, $v_2 = -0,01$, $v_3 = 0,24$, $v_4 = 0,55$. Jak widać dla ojców odległość pomiędzy kategoriami wykształcenia *średnie* i *wyższe* jest większa niż pozostałe, natomiast dla respondentów największa jest odległość pomiędzy dwiema najniższymi kategoriami. Parametr interakcji dla tego modelu wynosi $\delta = 68,71$. Przypomnijmy, że wartości te pozwalają zrekonstruować dowolny stosunek szans. Na przykład

$$\Theta_{31}^{XY} = 68,7^{(0,71-0,19)(-0,01-(-0,79))} = 5,56$$

Oznacza to, że proporcja osób z wykształceniem podstawowym do osób z wykształceniem niepełnym średnim jest ponad 5,5 razy większa wśród respondentów, których ojcowie mieli wykształcenie średnie aniżeli wśród respondentów których ojcowie mieli wykształcenie wyższe. Stosunek ten byłby mniejszy gdybyśmy porównywali proporcję dla innych sąsiednich kategorii wykształcenia respondenta bądź porównywali inne sąsiednie kategorie wykształcenia ojca, co wynika z tego, że różnice pomiędzy parametrami dla tych kategorii są mniejsze.

Jeśli chodzi o dane panelowe parametry dla kolejnych kategorii zmiennej opisującej sytuację materialną gospodarstw w 2000 roku wynoszą $u_1 = -0,61$ dla kategorii „z wielką trudnością", a dla kolejnych kategorii $u_2 = -0,27$, $u_3 = -0,07$, $u_4 = 0,28$,$u_5 = 0,68$. Jeśli chodzi o odpowiednie parametry dla kategorii z 2005 roku to wynoszą one

$v_1 = -0,64$, $v_2 = -0,28$, $v_3 = 0$, $v_4 = 0,28$,$v_5 = 0,65$. Jak widać, skalowanie to jest dość zbliżone dla 2000 i 2005 roku.

W odniesieniu do tablic o takich samych kategoriach zmiennej wierszowej i kolumnowej możliwe jest przyjęcie dodatkowego założenia, zgodnie z którym skalowanie zmiennej wierszowej i kolumnowej jest takie samo, tj. $u_i = v_i$ dla każdej wartości $i$ zmiennej $X$ oraz $Y$. Wówczas:

$$\Theta_{ij}^{XY} = \Theta_{ji}^{XY} = \delta^{(u_{(i+1)}-u_i)(u_{(j+1)}-u_j)} \tag{3.27}$$

Po przyjęciu tego założenia model wierszowo–kolumnowy może być postrzegany jako model quasi–symetrii. Będzie on nazywany symetrycznym logarytmiczno–multiplikatywnym modelem wierszowo–kolumnowym i oznaczany $(R = C)2$.

Warto zauważyć, że przyjęcie założenia o identycznym skalowaniu wartości dla zmiennej wierszowej i kolumnowej może być w odniesieniu do tablic ruchliwości i danych panelowych uzasadnione. Można oczekiwać, że odległości pomiędzy kolejnymi kategoriami wykształcenia dla ojca i syna są takie same, np. różnica pomiędzy wykształceniem podstawowym i średnim jest taka sama dla pokolenia synów i ojców. Podobnie można oczekiwać, że w przypadku badań panelowych pomiędzy kolejnymi badaniami nie wydarzyło się nic co uzasadniałoby przyjęcie odrębnego skalowania zmiennej porządkowej (lub nominalnej) dla porównywanych punktów czasowych. Oczywiście, jeśli istnieją przesłanki ku temu, że takie zmiany nastąpiły i na przykład w wyniku reform szkolnictwa kategorie wykształcenia syna i ojca w pewnym stopniu nie odpowiadają sobie, wówczas przyjęcie założenia o symetrycznym skalowaniu staje się kontrowersyjne.

Tak jak w poprzednich modelach parametry interakcji są symetryczne, jeśli kategoriami odniesienia są odpowiadające sobie kategorie $x_a$ oraz $y_a$ obydwu zmiennych:

$$\pi_{ij}^{XY} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \delta^{(u_i-u_a)(u_j-u_b)} \tag{3.28}$$

Przypomnijmy, że na parametry skalujące można nałożyć odpowiednie założenia, takie jak w formule 2.47. W stosunku do modelu niezależności mamy dodatkowo jeden parametr interakcji i $r-2$ niezależnych parametrów skalujących. Liczba stopni swobody tego modelu wynosi więc:

$$df = (r-1)^2 - 1 - (r-2) = (r-2)(r-1).$$

Model wierszowo–kolumnowy w postaci symetrycznej jest akceptowalny dla danych panelowych, co pokazuje wynik weryfikacji dla tabeli 3.24. Porównanie modelu w wersji asymetrycznej i symetrycznej tj. RC2 i (R=C)2 za pomocą testu warunkowego pokazuje, że przyjęcie identycznego skalowania dla zmiennej wierszowej i kolumnowej nie pogarsza dopasowania w sposób istotny statystycznie: $G^2 = 2,36$, $df = 3$

Tabela 3.33: Ilustracja symetrycznego modelu wierszowo–kolumnowy II typu — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\delta^{u_2u_2}$ | $\delta^{u_2u_3}$ | $\delta^{u_2u_4}$ | $\delta^{u_2u_5}$ |
| $x_3$ | 1 | $\delta^{u_3u_2}$ | $\delta^{u_3u_3}$ | $\delta^{u_3u_4}$ | $\delta^{u_3u_5}$ |
| $x_4$ | 1 | $\delta^{u_4u_2}$ | $\delta^{u_4u_3}$ | $\delta^{u_4u_4}$ | $\delta^{u_4u_5}$ |
| $x_5$ | 1 | $\delta^{u_5u_2}$ | $\delta^{u_5u_3}$ | $\delta^{u_5u_4}$ | $\delta^{u_5u_5}$ |

($p = 0,5$). Odpowiednie parametry tego modelu wynoszą: $u_1 = -1.53$, $u_2 = -0.67$, $u_3 = -0.07$, $u_4 = 0,68$,$u_5 = 1,59$. Należy zauważyć że odległości pomiędzy kolejnymi kategoriami są podobne, przy czym największe są odległości pomiędzy kategorią pierwszą i drugą jak również czwartą i piątą.

Model $(R = C)2$ nie wydaje się natomiast realistyczny w odniesieniu do tablicy ruchliwości. Test warunkowy pokazuje, że dopasowanie tego modelu jest istotnie gorsze niż modelu w postaci asymetrycznej: $G^2 = 15,7$, $df = 2$ ($p = 0,0004$). Wynik ten stanowi potwierdzenie dla wartości uzyskanych w asymetrycznej wersji tego modelu. Okazuje się, że uzyskane rozbieżności pomiędzy odległościami wyznaczonymi dla kategorii wykształcenia ojca i syna nie muszą być przypadkowe.

Zarówno w symetrycznej jak i asymetrycznej wersji modelu możliwe jest uwzględnienie specyfiki głównej przekątnej. Wiąże się to w jednym i w drugim przypadku — podobnie jak w poprzednio omawianych modelach — z dodaniem $r$ parametrów, po jednym dla każdej komórki przekątnej. Skupimy się na modelu symetrycznym. Zmodyfikowany w ten sposób model będzie oznaczany $Q(R = C)2$. Dopasowanie takiego modelu do danych panelowych jest bardzo dobre. Odpowiednie parametry przekątnej wynoszą $q_1 = 3,22$, $q_2 = 0,62$, $q_3 = 1,50$, $q_4 = 0,34$, $q_5 = 9,42$. Pamiętać należy, że stanowią one modyfikację modelu wierszowo–kolumnowego, ich interpretacja jest więc bardziej skomplikowana niż w przypadku modelu quasi–niezależności. Dodajmy, że podobnie jak w przypadku hipotezy $Q(R = C)1$, ma ona zastosowanie do tabeli o wymiarach co najmniej 5 na 5. W przeciwnym razie model ten, który jest szczególnym przypadkiem quasi–symetrii ma w stosunku do niego więcej stopni swobody, co jest sprzeczne.

Powyższe przykłady empiryczne dotyczą zmiennych porządkowych, warto jednak zauważyć, że — w odróżnieniu od modelu jednakowej interakcji i modelu wierszowo-kolumnowego I typu — omawiane modele można zastosować również wówczas, gdy

mamy do czynienia ze zmiennymi nominalnymi o takich samych kategoriach, przykładowo odnieść je do tablicy ruchliwości zawodowej. Dla tablicy 3.2 model $QRC2$ jest akceptowalny na poziomie istotności, $0,05$ ($G^2 = 15,9$, $df = 10$ $p = 0,1028$), natomiast model $Q(R = C)2$ na poziomie istotności $0,01$, tj. ($G^2 = 25,6$, $df = 14$, $p = 0,0288$). Porównanie obydwu modeli za pomocą testu warunkowego pokazuje, że hipotezę identycznego skalowaniu dla zawodu ojca i syna, należy odrzucić na poziomie istotności $0,05$, choć nie ma do tego podstaw na niższym poziomie $0,01$ ($G^2 = 10,3$, $df = 4$, $p = 0,0357$). Wyniki homogenicznego skalowania wynoszą $u_1 = -0,66$ dla inteligencji i wyższych kadr, $u_2 = -0,33$ dla pracowników umysłowych, $u_3 = -0,09$ dla prywatnych przedsiębiorców, $u_4 = 0,19$ dla robotników wykwalifikowanych, $u_5 = 0,42$ dla robotników niewykwalifikowanych, $u_6 = 0,47$ dla właścicieli gospodarstw i robotników rolnych. Wartości parametrów dla inteligencji i rolników różnią się najbardziej. Wynik ten wskazuje, że dla przedstawicieli tych kategorii w pokoleniu ojców, rozkłady wykształcenia syna różną się najbardziej, w tym sensie, że wartości odpowiednich stosunków szans odtworzone na podstawie modelu byłyby największe. Można powiedzieć również — na mocy homogeniczności parameramerametrów — że dwie powyższe kategorie w pokoleniu synów różnią się najbardziej jeśli chodzi o pochodzenie.

Do nieco innych wniosków prowadzą parametry modelu heterogenicznego: $u_1 = -0,69$, $u_2 = -0,32$, $u_3 = -0,04$, $u_4 = 0,20$, $u_5 = 0,53$, $u_6 = 0,79$ dla zmiennej wierszowej oraz $v_1 = -0,68$, $v_2 = -0,22$, $v_3 = -0,08$, $v_4 = 0,32$, $v_5 = 0,61$, $v_6 = 0,03$ dla zmiennej kolumnowej. O ile wartości parametrów dla przynależności społeczno–zawodowej ojca są podobne jak w poprzednim modelu, to parametr dla kategorii rolników w pokoleniu synów lokuje tę kategorię pośrodku skali, natomiast największy dystans jest pomiędzy robotnikami niewykwalifikowanymi a inteligencją. Wynik ten w pewnym sensie potwierdza, że nie należy formułować zbyt mocnych założeń co do uporządkowania kategorii tej zmiennej, gdyż jest to zmienna nominalna[24].

### 3.2.4 Modele dystansu i przekraczania barier

Jak zostało zasygnalizowane wcześniej modele jednakowej interakcji i wierszowo–kolumnowe stosowane są do opisu zależności między dowolnymi zmiennymi, a przyjęcie dodatkowych założeń może czynić je szczególnie użytecznymi w odniesieniu do tablic ruchliwości i danych panelowych. W tej części zaprezentowane zostaną modele, które są specyficzne dla tabel o takich samych kategoriach zmiennej wierszowej

[24] Oczywiście monotoniczność parametrów nie może stanowić dowodu na to, że zmienna jest mierzona na skali porządkowej, a jej brak, że jest mierzona na skali nominalnej. Wynik taki może stanowić jedynie sugestię, że mamy do czynienia z cechą wielowymiarową.

i kolumnowej, mówiąc inaczej, trudno byłoby je stosować o odniesieniu do danych innego typu W modelach tych wykorzystuje się informacje o porządkowym charakterze zmiennych. Jak zobaczymy, w poszczególnych modelach czyni się dodatkowe założenia, np. istotna będzie informacja o tym, o ile kategorii różni się sytuacja danej osoby (obiektu obserwacji) ze względu na dwie porównywane zmienne. Prezentując poszczególne modele wskazywane będą dodatkowe założenia pomiarowe, jakie są w nich *implicite* zawarte.

### Model zmiennego dystansu D

Analizując ruchliwość społeczną można oczekiwać, że prawdopodobieństwo zmiany pozycji społecznej syna w stosunku do pozycji ojca zależy od tego, jakiego rodzaju jest to zmiana. W przypadku gdy rozpatrujemy zmienną porządkową, na przykład wykształcenie, można wysunąć hipotezę, że przy kontroli rozkładów brzegowych zmiana pozycji o kilka poziomów jest relatywnie mniej prawdopodobna niż zmiana pozycji o jeden poziom, bądź odziedziczenie pozycji po ojcu. Co więcej, można rozróżnić zmiany pozycji „w dół" i „w górę", tj. awans bądź degradację pozycji syna w stosunku do pozycji ojca. Podobna hipoteza może mieć zastosowanie w odniesieniu do danych panelowych. Można jednak przypuszczać, że jeśli następuje zmiana, to jej prawdopodobieństwo zależy od jej kierunku oraz tego, ilu kategorii dotyczy.

Powyższe intuicje w sposób bardziej sprecyzowany sformułowane zostały w modelu zmiennego dystansu, który będzie oznaczany jako D (Goodman, 1972). Hipotezę tę można sformułować w podobny sposób, jak hipotezę o quasi–niezależności. Przypuśćmy, że interesuje nas rozkład łączny zmiennych $X$ oraz $Y$, czyli prawdopodobieństwa rozkładu łącznego $\pi_{ij}^{XY}$. Dla celów prezentacji modelu przyjmijmy, że zmienna $X$ oznacza wykształcenie ojca, zmienna $Y$ – wykształcenie syna, choć hipoteza możemy dotyczyć innych zmiennych o takich samych kategoriach, na przykład danych panelowych. Przekątna tej tabeli opisuje osoby, których wykształcenie nie zmieniło się w stosunku do wykształcenia ojca, pseudo–przekątne poniżej przekątnej wskazują na osoby, których wykształcenie pogorszyło się o odpowiednio 1, 2,... $r-1$ kategorii, pseudo–przekątne powyżej przekątnej opisują poprawę wykształcenia syna względem wykształcenia ojca w analogiczny sposób. Wyznaczmy zmienną $Z$, której wartości opisują, z którą podzbiorowością mamy do czynienia, tj. $Z$, będzie przyjmowała wartość $k$, gdzie $k = j - i$, czyli porównujemy, ile kategorii dzieli $i-$tą wartość zmiennej $X$ oraz $j-$tą wartość zmiennej $Y$. Przykładowo, $Z = -2$ dla tabeli 5 x 5 wskazuje na prawdopodobieństwa $\pi_{31}^{XY}$, $\pi_{42}^{XY}$, $\pi_{53}^{XY}$, czyli sytuację, gdy wykształcenie badanej osoby jest niższe o dwie kategorie od wykształcenia jej ojca, $Z = 1$, osoby, których wykształcenie jest o jedną kategorię wyższe, a $Z = 0$ sytuację, gdy wykształ-

cenie ojca jest takie samo jak wykształcenie respondenta. Tak więc, opisane powyżej wartości zmiennej $Z$, wskazują której przekątnej lub pseudo–przekątnej dotyczy podzbiorowość. Można jednak przyjąć, że nie wyczerpują one całej podzbiorowości, tj. uwzględnić dodatkowo, że zmienna $Z$ przyjmuje wartość $Z = r$, która nie wskazuje na konkretną pseudo–przekątną.

Prawdopodobieństwa poszczególnych wartości zmiennej $Z$ oznaczone będą $P(Z = k) = \beta_k$, natomiast $P(Z = r) = \beta_r$. Określają one odsetki poszczególnych podzbiorowości. Prawdopodobieństwo rozkładu łącznego wynosi więc:

$$\pi_{ij}^{XY} = \sum_{k=-(r-1)}^{r-1} \pi_{ij(k)}^{XY(Z)} \beta_k + \pi_{ij(r)}^{XY(Z)} \beta_r, \tag{3.29}$$

a ponieważ podzbiorowości $Z = k$ wykluczają się wzajemnie, powyższe równanie można zapisać jako:

$$\pi_{ij}^{XY} = \pi_{ij(k)}^{XY(Z)} \beta_k + \pi_{ij(r)}^{XY(Z)} \beta_r. \tag{3.30}$$

Zgodnie z modelem dystansu, zachodzą następujące warunki opisujące prawdopodobieństwa w poszczególnych podzbiorowościach i relacje między nimi:

$$\pi_{ab(k)}^{XY(Z)} = 0, \text{ dla każdej pary } a, b, \text{ takiej, że } b - a \neq k, \text{ oraz} \tag{3.31}$$

$$\Theta_{ij(r)}^{XY(Z)} = 1, \text{ dla dowolnej pary } i, j, \text{ oraz} \tag{3.32}$$

$$\frac{\pi_{cd(k)}^{XY(Z)}}{\pi_{ef(k)}^{XY(Z)}} = \frac{\pi_{cd(r)}^{XY(Z)}}{\pi_{ef(r)}^{XY(Z)}}, \text{ dla wartości } c, d, e, f, \text{ takich, że } d - c = f - e = k, \tag{3.33}$$

przy czym warunki 3.31 oraz 3.33 zachodzą dla każdego $k$. Powyższe sformułowanie hipotezy wymaga krótkiego omówienia. Ich ilustracją jest tabela w których obydwie zmienne $X$ oraz $Y$ przyjmują 5 wartości, i przedstawiono wybrane podzbiorowości: $Z = 0$, $Z = 2$, oraz $Z = 5$. Warunek 3.31 głosi, że w każdej podzbiorowości $Z = k$ niezerowe prawdopodobieństwa występują jedynie na odpowiednich pseudo–przekątnych, tak więc wartości $Z = k$ opisują osoby homogeniczne, ze względu na to ile kategorii dzieli ich wykształcenie od wykształcenia ich ojca. W tabeli 3.34 ich przykładem są podzbiorowości $Z = 0$ oraz $Z = 2$. Zgodnie z warunkiem 3.32, w podzbiorowości $Z = r$ zmienne są niezależne stochastycznie, tj. wszystkie stosunki szans są równe 1. W naszym przykładzie jest to podzbiorowość $Z = 5$. Z dwóch pierwszych warunków wynika, że na odsetek w każdej komórce rozkładu łącznego zmiennych $X$ i $Y$ składają się z jednej strony osoby, których wykształcenie nie zależy od wykształcenia ojca ($Z = r$), jak również te, których wykształcenie różni się o konkretną liczbę kategorii ($Z = k$). Przykładowo, prawdopodobieństwo $\pi_{24}^{XY}$ jest równe:

$$\pi_{24}^{XY} = \pi_{24(2)}^{XY(Z)} \beta_2 + \pi_{24(5)}^{XY(5)} \beta_5.$$

Ostatni warunek 3.33 pokazuje, że relacje pomiędzy odpowiednimi niezerowymi prawdopodobieństwami w każdej podzbiorowości $(Z = k)$, są takie jak w podzbiorowości $(Z = r)$. Zauważmy, że w tabeli 3.34 zachodzi:

$$\frac{\pi^{XY(Z)}_{24(2)}}{\pi^{XY(Z)}_{35(2)}} = \frac{\pi^{XY(Z)}_{24(5)}}{\pi^{XY(Z)}_{35(5)}} = \frac{a^X_2 a^Y_4}{a^X_3 a^Y_5}.$$

Podobnie:

$$\frac{\pi^{XY(Z)}_{22(0)}}{\pi^{XY(Z)}_{44(0)}} = \frac{\pi^{XY(Z)}_{22(5)}}{\pi^{XY(Z)}_{44(5)}} = \frac{a^X_2 a^Y_2}{a^X_4 a^Y_4}.$$

Tabela 3.34: Ilustracja modelu zmiennego dystansu, uwzględniająca dekompozycję rozkładu w całej zbiorowości na podzbiorowości wyróżnione ze względu na zmienną $Z$

| $Z = 0$ | | | | | |
|---|---|---|---|---|---|
| $X \backslash Y$ | 1 | 2 | 3 | 4 | 5 |
| 1 | $a^X_1 a^Y_1$ | 0 | 0 | 0 | 0 |
| 2 | 0 | $a^X_2 a^Y_2$ | 0 | 0 | 0 |
| 3 | 0 | 0 | $a^X_3 a^Y_3$ | 0 | 0 |
| 4 | 0 | 0 | 0 | $a^X_4 a^Y_4$ | 0 |
| 5 | 0 | 0 | 0 | 0 | $a^X_5 a^Y_5$ |
| $Z = 2$ | | | | | |
| $X \backslash Y$ | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | 0 | $a^X_1 a^Y_3$ | 0 | 0 |
| 2 | 0 | 0 | 0 | $a^X_2 a^Y_4$ | 0 |
| 3 | 0 | 0 | 0 | 0 | $a^X_3 a^Y_5$ |
| 4 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 |
| $Z = 5$ | | | | | |
| $X \backslash Y$ | 1 | 2 | 3 | 4 | 5 |
| 1 | $a^X_1 a^Y_1$ | $a^X_1 a^Y_2$ | $a^X_1 a^Y_3$ | $a^X_1 a^Y_4$ | $a^X_1 a^Y_5$ |
| 2 | $a^X_2 a^Y_1$ | $a^X_2 a^Y_2$ | $a^X_2 a^Y_3$ | $a^X_2 a^Y_4$ | $a^X_2 a^Y_5$ |
| 3 | $a^X_3 a^Y_1$ | $a^X_3 a^Y_2$ | $a^X_3 a^Y_3$ | $a^X_3 a^Y_4$ | $a^X_3 a^Y_5$ |
| 4 | $a^X_4 a^Y_1$ | $a^X_4 a^Y_2$ | $a^X_4 a^Y_3$ | $a^X_4 a^Y_4$ | $a^X_4 a^Y_5$ |
| 5 | $a^X_5 a^Y_1$ | $a^X_5 a^Y_2$ | $a^X_5 a^Y_3$ | $a^X_5 a^Y_4$ | $a^X_5 a^Y_5$ |

Ilustrację modelu zmiennego dystansu w formie parametrycznej przedstawia również tabela 3.35. Ilustracja ta jest zgodna z następującą parametryzacją:

$$\pi_{ij}^{XY} = d \cdot d_i^X \cdot d_j^Y \cdot s_k \qquad \text{gdzie } k = j - i \tag{3.34}$$

Zgodnie z tym modelem parametr interakcji zależy od tego, o ile komórek rozpatrywana kombinacja dwóch zmiennych oddalona jest od głównej przekątnej i w którą stronę następuje przesunięcie.

Zauważmy, że komórki na kolejnych pseudo–przekątnych, posiadają ten sam parametr interakcji, np. prawdopodobieństwa $\pi_{13}^{XY}$, $\pi_{24}^{XY}$, $\pi_{35}^{XY}$, charakteryzuje ten sam parametr $s_2$. W odniesieniu do tablicy ruchliwości parametr $s_2$ charakteryzuje sytuację „awansu" o dwie kategorie wykształcenia. Analogicznie parametry $s_{-2}$ charakteryzują sytuację „degradacji" o dwie kategorie.

Tabela 3.35: Ilustracja modelu zmiennego dystansu D — parametry interakcji

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | $s_1$ | $s_2$ | $s_3$ | $s_4$ |
| $x_2$ | $s_{(-1)}$ | 1 | $s_1$ | $s_2$ | $s_3$ |
| $x_3$ | $s_{(-2)}$ | $s_{(-1)}$ | 1 | $s_1$ | $s_2$ |
| $x_4$ | $s_{(-3)}$ | $s_{(-2)}$ | $s_{(-1)}$ | 1 | $s_1$ |
| $x_5$ | $s_{(-4)}$ | $s_{(-3)}$ | $s_{(-2)}$ | $s_{(-1)}$ | 1 |

Porównując dwie ilustracje modelu zmiennego dystansu — z tabel 3.34 oraz 3.35 — widać, że na wielkości parametrów z drugiej tabeli wpływ mają wielkości poszczególnych podzbiorowości wyróżnionych ze względu na zmienną $Z$. Im mniejsza jest dana podzbiorowość $Z = k$, w stosunku do innych podzbiorowości, tym relatywnie mniejsza jest wartość parametru $s_k$. W tabeli przyjęto, że wielkość $s_0$ jest równa 1, dlatego relatywną częstość występowanie kolejnych pseudo–przekątnych odnosimy do głównej przekątnej. Przykładowo, wielkość $s_2$ wskazuje, ile razy bardziej (mniej) jest prawdopodobne — przy kontroli różnic w rozkładach brzegowych — że gdy zmienna $Y$ ma wartość większą o dwie kategorie od zmiennej $X$, w stosunku do tego, ze obydwie zmienne przyjmują tę samą wartość. Warto podkreślić, że w modelu tym — podobnie jak prezentowanych wcześniej modelach QN i QhN — nie odpowiadamy na pytanie o wielkość podzbiorowości wyróżnionych ze względu na zmienną $Z$. Mówiąc inaczej, ten sam rozkład zgodny z modelem dystansu „zdekomponować" w różny sposób.

Zauważmy, że w modelu tym ukryte jest dodatkowe założenie pomiarowe w stosunku do rozpatrywanych wcześniej modeli. Zakładamy nie tylko uporządkowanie kategorii obydwu analizowanych zmiennych, ale również, że „wielkość" zmiany pozy-

cji (awansu, degradacji) możemy definiować za pomocą różnicy liczby kategorii dwóch porównywanych zmiennych. Założenie to wydaje się dość kontrowersyjne. Co prawda można argumentować, że testując model założenie to również podlega weryfikacji, tj. jeśli model jest dobrze dopasowany do danych, można przypuszczać, że różnica w liczbie kategorii dobrze zdaje sprawę z „natężenia" zmiany. Trzeba jednak zauważyć, że dobre dopasowanie do danych, może być w jakiejś mierze „przypadkowe", tj. badane zjawisko np. ruchliwość społeczna powinna być opisywana przez zupełnie inny proces (inny model) ale obydwa dają podobny rozkład oczekiwany.

Hipotezę powyższą można sformułować również za pomocą lokalnych stosunków szans dotyczących komórek o określonej różnicy pomiędzy wartościami obydwu zmiennych. Zgodnie z modelem zmiennego dystansu zachodzi:

$$\Theta^{XY}_{i(i+k)} = \frac{\pi^{XY}_{i(i+k)} / \pi^{XY}_{i(i+k+1)}}{\pi^{XY}_{(i+1)(i+k)} / \pi^{XY}_{(i+1)(i+k+1)}} = \Theta_k \tag{3.35}$$

przy czym $i = 1, \ldots r-1$, $k = -(r-2), \ldots, (r-2)$ oraz $1 \leq i + k \leq r - 1$. Powyższy warunek określa, że stosunki szans są specyficzne dla tych kombinacji dwóch zmiennych, które są oddalone o $k$ kategorii od przekątnej, tak więc równe sobie są np. stosunki szans $\Theta^{XY}_{13}$, $\Theta^{XY}_{24}$, podobnie równe sobie są stosunki szans $\Theta^{XY}_{21}$, $\Theta^{XY}_{32}$, $\Theta^{XY}_{43}$. itd. W liczniku zdefiniowanego powyżej stosunku szans porównujemy prawdopodobieństwo zmiany pozycji o $k$-kategorii w stosunku do zmiany o $k+1$ kategorii. W mianowniku porównujemy zmianę pozycji o $k-1$ kategorii w stosunku do zmiany o $k$ kategorii. Im wyższa jest wartość zdefiniowanego powyżej stosunku szans, tym bardziej zmiana o $k$ kategorii jest prawdopodobna relatywnie do zmiany pozycji o $k-1$ i $k+1$ kategorii. Zauważmy, że jeśli w powyższej formule $k = 0$, tj. nie mamy do czynienia ze zmianą pozycji, wówczas zgodnie z tym warunkiem, lokalne stosunki szans wyznaczone dla komórek na głównej przekątnej są sobie równe, tj. $\Theta^{XY}_{11} = \Theta^{XY}_{22}$, itd.

Z warunku powyższego wynika, że specyficzne są stosunki szans porównujące komórki umieszczone na głównej przekątnej z komórkami oddalonymi od niej o $m$ kategorii, tj.

$$\Theta^{X\ \ Y}_{i/(i+m);i/(i+m)} = \frac{\pi^{XY}_{ii} / \pi^{XY}_{i(i+m)}}{\pi^{XY}_{(i+m)i} / \pi^{XY}_{(i+m)(i+m)}} = \Theta_m \tag{3.36}$$

Powyższy stosunek szans porównuje proporcję liczby osób, które dziedziczą pozycję $i$ do liczby osób, które z pozycji $i$ awansowały o $m$ kategorii z proporcją liczby osób, które zostały „zdegradowane" o $m$ kategorii z pozycji $i+m$ do liczby osób które pozostały na tej pozycji. Wielkość ta zależy tylko od tego, ilu kategorii $m$ ten awans bądź degradacja dotyczyły.

Tabela 3.36: Wyniki weryfikacji modeli dystansu i przekraczania barier dla tabel 3.5, 3.6

| Model | df | $\chi^2$ | $G^2$ | $\Delta$ |
|---|---|---|---|---|
| Wyniki dla tabeli 3.5 (tablica ruchliwości edukacyjnej) | | | | |
| D | 4 | 11,0 ($p = 0,0265$) | 11,4 ($p = 0,0226$) | 2,4 |
| QD | 1 | 5,6 ($p = 0,0177$) | 6,0 ($p = 0,0147$) | 0,9 |
| DS | 6 | 12,7 ($p = 0,0487$) | 13,7 ($p = 0,0331$) | 2,4 |
| QDS | 3 | 7,4 ($p = 0,0591$) | 8,1 ($p = 0,0439$) | 1,1 |
| CP | 6 | 36,9 ($p < 0,0001$) | 38,0 ($p < 0,0001$) | 4,5 |
| QCP | 4 | 7,7 ($p = 0,1032$) | 8,3 ($p = 0,0825$) | 1 |
| FD | 8 | 37,2 ($p < 0,0001$) | 38,7 ($p < 0,0001$) | 4,7 |
| QFD | 4 | 7,7 ($p = 0,1032$) | 8,3 ($p = 0,0825$) | 1,0 |
| $FD^{1,5}$ | 8 | 12,8 ($p = 0,1204$) | 13,8 ($p = 0,0862$) | 2,4 |
| $QFD^{1,5}$ | 4 | 7,8 ($p = 0,0997$) | 8,7 ($p = 0,0697$) | 1,2 |
| Wyniki dla tabeli 3.6 (dane panelowe) | | | | |
| D | 9 | 14,4 ($p = 0,1080$) | 14,4 ($p = 0,1081$) | 2,9 |
| QD | 5 | 5,4 ($p = 0,3690$) | 5,5 ($p = 0,3530$) | 1,3 |
| DS | 12 | 15,0 ($p = 0,2430$) | 15,0 ($p = 0,2392$) | 3,1 |
| QDS | 8 | 6,0 ($p = 0,6429$) | 6,2 ($p = 0,6295$) | 1,3 |
| CP | 12 | 45,9 ($p < 0,0001$) | 47,0 ($p < 0,0001$) | 6,4 |
| QCP | 9 | 2,9 ($p = 0,9699$) | 2,9 ($p = 0,9698$) | 0,9 |
| FD | 15 | 54,9 ($p < 0,0001$) | 56,4 ($p < 0,0001$) | 6,6 |
| QFD | 10 | 7,3 ($p = 0,6929$) | 7,4 ($p = 0,6907$) | 1,5 |
| $FD^{1,5}$ | 15 | 17,3 ($p = 0,3012$) | 17,5 ($p = 0,2898$) | 3,2 |
| $QFD^{1,5}$ | 10 | 6,2 ($p = 0,7990$) | 6,3 ($p = 0,7875$) | 1,2 |

Formułę na liczbę stopni swobody można wyznaczyć, pamiętając, że jest ona równa liczbie niezależnych ograniczeń, jakie nakładamy na lokalne stosunki szans. Hipoteza 3.35 zakłada, że lokalne stosunki szans na głównej przekątnej, tj. $\Theta_{ii}^{XY}$ są sobie równe. Ich liczba wynosi $r-1$, tak więc liczba warunków wynosi $r-2$. Kolejno, mamy dwie pseudo-przekątne dotyczące komórek, które przylegają do głównej przekątnej, więc liczba tych ograniczeń wynosi $2(r-3)$. Analogicznie można rozpatrywać kolejne pseudo–przekątne. Liczba stopni swobody dana jest więc wzorem:

$$df = (r-2) + 2[(r-3) + (r-4) + \ldots + 1] = (r-2) + (r-3)(r-2) = (r-2)^2.$$

Tak więc model D w stosunku do niezależności stochastycznej posiada $2r-3$ niezależnych parametrów więcej. Dla tabeli 3.35 daje nam to 7 niezależnych parametrów, natomiast w tabeli oznaczono — dla celów ilustracyjnych — 8 parametrów $s_k$. Zauważmy jednak, że można przyjąć, że $s_{(-4)}=1$, bądź $s_4=1$ i nie zmieni to ogólności prezentowanej hipotezy, bowiem, przyjmując jeden z tych warunków nie zakładamy nic dodatkowo w odniesieniu do stosunków szans, które komórki te definiują[25].

Tabela 3.37: Rozkład oczekiwany zgodny z modelem D dla danych z tabeli 3.6

| Czy przy aktualnym dochodzie netto Pana(i) gospodarstwo domowe wiąże koniec z końcem? | | | | | | |
|---|---|---|---|---|---|---|
| Odpowiedzi w 2000 roku ($X$) | Odpowiedzi w 2005 roku ($Y$) | | | | | Suma |
| | 1 | 2 | 3 | 4 | 5 | |
| 1. Z wielką trudnością | 265,0 | 177,0 | 112,4 | 24,9 | 2,9 | 582,2 |
| 2. Z trudnością | 109,0 | 145,1 | 158,2 | 38,1 | 6,0 | 456,4 |
| 3. Z pewną trudnością | 56,0 | 115,8 | 251,8 | 104,2 | 18,0 | 545,8 |
| 4. Raczej łatwo | 8,5 | 25,7 | 86,9 | 71,7 | 21,3 | 214,1 |
| 5. Łatwo | 1,2 | 2,2 | 11,1 | 14,2 | 8,4 | 37,1 |
| Suma | 439,7 | 465,8 | 620,4 | 253,1 | 56,6 | 1835,6 |

Model ten jest akceptowalny dla danych ruchliwości edukacyjnej na poziomie istotności 0,01, natomiast dla danych panelowych można go zaakceptować nawet na na wyższym poziomie istotności. Rozkład oczekiwany przedstawiony został w tabeli 3.37. Aby oszacować model zgodnie z metodą największej wiarygodności należy założyć, że rozkłady brzegowe są szacowane na podstawie próby, jak również dla każdego, k, takiego, że $k=j-i$ spełniony jest warunek:

$$\sum_{i}^{r}\sum_{j}^{r}\alpha_k\cdot\widehat{\pi}_{ij}^{XY}=\sum_{i}^{r}\sum_{j}^{r}\alpha_k\cdot p_{ij}^{XY} \tag{3.37}$$

gdzie $\alpha_k$ jest równe 1, dla takich komórek, że $j-i=k$ i 0 w przeciwnej sytuacji. Innymi słowy, suma oszacowanych prawdopodobieństw (liczebności) na poszczególnych pseudo–przekątnych jest równa analogicznej sumie częstości (liczebności) w próbie, np. dla tabeli 3.37 spełniony jest warunek:

$$\widehat{\pi}_{13}^{XY}+\widehat{\pi}_{24}^{XY}+\widehat{\pi}_{35}^{XY}=p_{31}^{XY}+p_{24}^{XY}+p_{35}^{XY}=168,5.$$

[25]Nie można założyć, że obydwa te ograniczenia są spełnione równocześnie, bo wynikałoby z tego, że $\Theta^{X\ \ Y}_{i/(i+m);i/(i+m)}=1$, co nie wynika z hipotezy 3.35

Parametry modelu D dla danych panelowych są następujące: $s_1 = 0,83$, $s_2 = 0,40$, $s_3 = 0,17$, $s_4 = 0,06$ dla komórek powyżej przekątnej oraz $s_{-1} = 0,60$, $s_{-2} = 0,23$, $s_{-3} = 0,06$, $s_{-4} = 0,03$ dla komórek poniżej przekątnej. Tak więc, że kontrolując różnice w rozkładach brzegowych sytuacja, w której ocena położenia finansowego gospodarstwa poprawiła się (w 2005 roku w stosunku do 2000 roku) o jedną kategorię jest 1,2 (tj. 1/0,83) razy mniej prawdopodobna, niż sytuacja, w której badany oceniał kondycję finansową identycznie. Parametry te pozwalają również odtworzyć dowolny stosunek szans, na przykład:

$$\Theta_{24}^{XY} = \frac{s_2/s_3}{s_1/s_2} = \frac{0,4/0,17}{0,83/0,4} = \frac{38,1 \cdot 18,0}{6,0 \cdot 104,2} = 1,1$$

Stosunki szans wyróżnione w ten sposób nie są symetryczne tj. wielkości $\Theta_{42}^{XY}$, $\Theta_{31}^{XY}$, są inne:

$$\Theta_{42}^{XY} = \frac{s_{(-2)}/s_{(-3)}}{s_{(-1)}/s_{(-2)}} = \frac{0,23/0,06}{0,6/0,23} = 1,47.$$

Tak więc proporcje te różniłyby się w większym stopniu. Parametry wyznaczone dla tablicy ruchliwości wynoszą odpowiednio: $s_1 = 1,26$, $s_2 = 1,01$, $s_3 = 0,68$ oraz $s_{-1} = 0,31$, $s_{-2} = 0,07$, natomiast wartość parametru $s_{-3}$ jest bardzo mała i wynosi w przybliżeniu zero. Należy jednak zauważyć model D nie głosi nic na temat jedynej komórki związanej z tym parametrem, w konsekwencji jej liczebność oczekiwana pokrywa się z odpowiednią liczebnością w próbie , tj. $\widehat{F}_{41} = \widehat{f}_{41}$ (podobnie $\widehat{F}_{14} = \widehat{f}_{14}$). Biorąc pod uwagę, że komórka ta w próbie ma zerową liczebność wynik taki nie zaskakuje, ale też trudno na jego podstawie wyciągać daleko idące wnioski.

Model powyższy nie jest przykładem quasi–symetrii: stosunki szans w modelu D nie są symetryczne, nie są też symetryczne parametry $s_k$. Możliwe jest sformułowanie wersji symetrycznej modelu opisanego powyżej. Konieczne jest wówczas dodanie do warunków 3.35, definiujących model zmiennego dystansu dodanie założenia o symetrii stosunków szans (3.17). Tak więc, w modelu tym lokalne stosunki szans są równe:

$$\Theta_{i(i+k)}^{XY} = \frac{\pi_{i(i+k)}^{XY}/\pi_{i(i+k+1)}^{XY}}{\pi_{(i+1)(i+k)}^{XY}/\pi_{(i+1)(i+k+1)}^{XY}} = \Theta_k \tag{3.38}$$

$$\Theta_{ij}^{XY} = \Theta_{ji}^{XY} \tag{3.39}$$

przy czym $i = 1, \ldots r-1$, $j = 1, \ldots r-1$, $k = 0, 1, \ldots r-2$ oraz $(i+k) \leq r-1$. Oznacza to na przykład, że nie tylko zachodzi $\Theta_{13}^{XY} = \Theta_{24}^{XY}$, ale również $\Theta_{13}^{XY} = \Theta_{31}^{XY} = \Theta_{42}^{XY}$. Parametryzacja tego modelu jest podobna do formuły 3.34 przy czym $k = |j-i|$. Ilustracją tego modelu jest tabela 3.38. W stosunku do tabeli 3.35 parametry interakcji są symetryczne, tj. $s_{-i} = s_i$. Wielkość parametru zależy więc wyłącznie od tego jak daleko leży pseudo–przekątna od głównej przekątnej, nie ma natomiast znaczenia, po

której stronie przekątnej leży dana komórka, tj. „awans" bądź „degradacja" są relatywnie tak samo prawdopodobne (przy kontroli różnic w rozkładach brzegowych) i zależą jedynie od tego, ilu kategorii dotyczą.

Tabela 3.38: Ilustracja symetrycznego modelu dystansu DS

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | $s_1$ | $s_2$ | $s_3$ | $s_4$ |
| $x_2$ | $s_1$ | 1 | $s_1$ | $s_2$ | $s_3$ |
| $x_3$ | $s_2$ | $s_1$ | 1 | $s_1$ | $s_2$ |
| $x_4$ | $s_3$ | $s_2$ | $s_1$ | 1 | $s_1$ |
| $x_5$ | $s_4$ | $s_3$ | $s_2$ | $s_1$ | 1 |

Ponieważ symetryczna wersja modelu zmiennego dystansu jest prostsza niż postać asymetryczna, jego liczba stopni swobody jest większa. Tak jak poprzednio mamy $(r-2)$ ograniczeń dotyczących stosunków szans na głównej przekątnej, czyli $\Theta_{11}^{XY} = \Theta_{22}^{XY}$, itd. Ponadto, równe sobie są wszystkie stosunki *wyznaczone*[26] dla pseudo–przekątnej przylegającej go głównej przekątnej, tj. $\Theta_{12}^{XY} = \Theta_{21}^{XY} = \Theta_{23}^{XY} = \Theta_{32}^{XY}$, itd. Ograniczeń tych jest $2(r-2)-1$. Analogicznie rozpatrujemy kolejne pseudo–przekątne. Łączna liczba tego typu warunków określa liczbę stopni swobody modelu, tj:

$$df = (r-2) + [2(r-2)-1] + [2(r-3)-1] + \ldots + 1 = (r-2)(r-1).$$

Wynika z tego, że model posiada $r-1$ parametrów więcej niż model niezależności stochastycznej. W tabeli 3.38 są to parametry $s_k$ opisujące poszczególne pseudo–przekątne. Aby oszacować rozkład oczekiwany metodą największej wiarygodności rozkłady brzegowe są szacowane na podstawie danych z próby. Ponadto muszą być spełnione warunki podane w formule 3.37 dotyczące modelu D, przy czym $\alpha_k$ jest równe 1, dla takich komórek, że $k = |j-i|$ i 0 w przeciwnej sytuacji.

Model ten — mimo, że prostszy od asymetrycznej wersji — może być zaakceptowany w przypadku tablicy ruchliwości edukacyjnej na poziomie istotności 0,01. Test warunkowy pokazuje, że dodatkowe założenie dotyczące quasi–symetrii nie pogarsza dopasowania w sposób istotny statystycznie: $G^2 = 2,32$, $df = 2$ $(p = 0,31)$. W tabeli 3.39 przedstawiony jest rozkład oczekiwany zgodny z modelem DS dla tablicy 3.39. Parametry dla tego modelu wynoszą: $s_1 = 0,62$, $s_2 = 0,25$, $s_3 = 0,08$. Wszystkie

[26]Rozróżnienie pomiędzy lokalnym stosunkiem szans *wyznaczonym* dla danej komórki, bądź ją *obejmującym* została opisana pod tabelą 3.1.

stosunki szans w tej tablicy są symetryczne, na przykład:

$$\Theta_{12}^{XY} = \Theta_{21}^{XY} = s_1^2/s_2 = 0,62^2/0,25 = 1,53$$

co oznacza, że proporcja liczby osób, które „awansowały" o jedną kategorię w stosunku do pozycji ojca, do liczby osób które „awansowały" o dwie kategorie, jest ponad 1,5 razy większa niż proporcja liczby osób, które odziedziczyły pozycję po ojcu do liczby osób, które poprawiły swoje wykształcenie o jedną kategorię. Z symetrii powyższej wielkości wynika również, że proporcja liczby osób, których pozycja pogorszyła się jedną kategorię w stosunku do pozycji ojca, do liczby osób, których pozycja pogorszyła się o dwie kategorie, jest ponad 1,5 razy większa niż proporcja liczby osób, które odziedziczyły pozycję po ojcu do liczby osób, których sytuacja pogorszyła się o jedną kategorię.

W przypadku danych panelowych dopasowanie do danych modelu DS wydaje się zadowalające. Test warunkowy porównujący modele D i DS pokazuje, że przyjęcie założenia o symetrycznej interakcji, trudno byłoby zakwestionować $G^2 = 0,62$, $df = 3$ ($p = 0,89$). Odpowiednie parametry wynoszą $s_1 = 0,712$, $s_2 = 0,30$, $s_3 = 0,11$, $s_4 = 0,04$.

Tabela 3.39: Rozkład oczekiwany zgodny z modelem DS dla tablicy 3.5

| Wykształcenie ojca | Wykształcenie syna | | | | Suma |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 1. Podstawowe i niepełne podstawowe | 335,9 | 515,3 | 193,9 | 99,9 | 1145 |
| 2. Niepełne średnie (w tym zasadnicze zawodowe) | 30,6 | 119,3 | 70,4 | 44,7 | 265 |
| 3. Ukończone średnie | 7,4 | 45,3 | 67,8 | 67,5 | 188 |
| 4. Niepełne wyższe i wyższe | 1,1 | 8,1 | 19,0 | 47,9 | 76 |
| Suma | 375 | 688 | 351 | 260 | 1674 |

Obydwa modele zmiennego dystansu — w symetrycznej i asymetrycznej wersji — można zmodyfikować uwzględniając dodatkowo specyfikę komórek położonych na głównej przekątnej, będziemy oznaczać je odpowiednio QDS i QD. Oznacza to, że warunek 3.33 nie dotyczy głównej przekątnej a jedynie poszczególnych pseudo–przekątnych. Mówiąc inaczej, relacje pomiędzy poszczególnymi prawdopodobieństwami na głównej przekątnej mogą być specyficzne, niekoniecznie takie same jak dla podzbiorowości opisywanej za pomocą niezależności (czyli $Z = r$).

Hipotezę asymetryczną oznaczymy jako QD i ilustruje ją tabela 3.40. Symetryczną wersję oznaczymy jako QDS. Uwzględnienie specyfiki głównej przekątnej wymaga w obydwu przypadkach uwzględnienia dodatkowo $r-1$ niezależnych parametrów. Mówiąc inaczej, można założyć, że jeden z parametrów 3.40 dotyczących komórek na głównej przekątnej jest równy 1. Liczba stopni swobody wynosi dla modelu QD:

$$df = r^2 - 5r + 5,$$

oraz

$$df = (r-1)(r-3)$$

dla modelu QDS. W Aneksie zamieszczona została formuła dotycząca obydwu modeli sformułowana za pomocą lokalnych stosunków szans, jak również bardziej precyzyjne uzasadnienie liczby stopni swobody.

Tabela 3.40: Ilustracja modelu dystansu QD — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $q_1$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ |
| $x_2$ | $s_{(-1)}$ | $q_2$ | $s_1$ | $s_2$ | $s_3$ |
| $x_3$ | $s_{(-2)}$ | $s_{(-1)}$ | $q_3$ | $s_1$ | $s_2$ |
| $x_4$ | $s_{(-3)}$ | $s_{(-2)}$ | $s_{(-1)}$ | $q_4$ | $s_1$ |
| $x_5$ | $s_{(-4)}$ | $s_{(-3)}$ | $s_{(-2)}$ | $s_{(-1)}$ | $q_5$ |

Jak pokazuje tabela 3.36 dopasowanie modeli QD i QDS dla danych panelowych jest dobre, natomiast dla tablicy ruchliwości edukacyjnej modele takie są akceptowalne na poziomie istotności 0,01. Modele D oraz QD są zagnieżdżone odpowiednio w modelach QD i QDS. Odpowiednie testy warunkowe pokazują na ile dodanie parametrów głównej przekątnej poprawia dopasowanie do danych. Okazuje się, że choć następuje poprawa w dopasowaniu do danych nie jest ona istotna statystycznie. Na przykład, w odniesieniu do danych panelowych $G^2 = 15,0 - 6,2 = 8,88$, $df = 4$ ($p = 0,06$). Wynika to z tego, że modele D i DS mimo, że nie uwzględniają parametrów głównej przekątnej, to w pewnym sensie uwzględniają specyfikę komórek na niej położonych. Mówiąc dokładniej, uwzględniają one tendencję do relatywnie częstszego (rzadszego) dziedziczenia pozycji bądź udzielania tej samej odpowiedzi w dwóch punktach czasowych. Poszczególne pseudo–przekątne, są opisane przez parametry interakcji i jeśli są one np. mniejsze od 1, wskazuje to —na zasadzie kontrastu — na relatywnie częstsze występowanie komórek głównej przekątnej.

Z drugiej strony modele D i DS nie różnicują pod tym względem poszczególnych komórek, tj. zakładają, że tendencja do występowania tej samej kategorii zmiennej wierszowej i kolumnowej dotyczy poszczególnych kategorii w takim samym stopniu, na przykład tendencja do dziedziczenia poziomu wykształcenia jest taka sama dla poszczególnych kategorii. W tym sensie modele D i DS podobne są do omawianego wcześniej modelu QhN (model ten jest zagnieżdżony zarówno w modelu D jak i modelu DS). Z tych samych względów uwzględnienie głównej przekątnej w obydwu modelach D i QD wymaga $(r-1)$ a nie $r$ dodatkowych parametrów.

Na koniec warto zauważyć, że dopasowanie modelu QDS i QS jest identyczne dla tablicy ruchliwości edukacyjnej. Wynik ten nie jest przypadkowy. Jeśli model QDS formułujemy dla tabeli o wymiarach 4 na 4 wówczas, obydwa modele są sobie równoważne. Uzasadnienie zostało zamieszczone w Aneksie.

### Model przekraczania barier CP

Opisywany w poprzedniej części model zmiennego dystansu przy opisie zmian pozycji społecznej w tablicy ruchliwości, bądź zmianie wartości porównywanej zmiennej w przypadku danych panelowych uwzględniał jedynie różnicę w liczbie kategorii pomiędzy porównywaną zmienną wierszową i kolumnową, nie uwzględniał natomiast, które kategorie ze sobą porównujemy. W rzeczywistości założenie to nie musi być spełnione: przykładowo, w odniesieniu do wykształcenia „dystans" pomiędzy kategoriami *podstawowe — niepełne średnie* może być inny niż „dystans" pomiędzy kategoriami *niepełne średnie—średnie*, co w tym modelu nie jest uwzględniane. Specyfikę dystansów pomiędzy kolejnymi kategoriami obydwu zmiennych uwzględnia sformułowany przez Goodmana (1972*c*) model *przekraczania barier* (*crossing parameter model*), który oznaczany będzie jako CP. Jeśli kolejnymi kategoriami wykształcenia są podstawowe, niepełne średnie, średnie to w modelu tym inaczej traktowana jest zmiana wykształcenia na nieukończone średnie w stosunku do podstawowego wykształcenia ojca niż zmiana pozycji na wykształcenie średnie w stosunku do wykształcenia niepełnego średniego. Istotne jest więc nie tyle *ilu*, ale *jakich* kategorii dotyczyła zmiana.

Zgodnie z tą hipotezą wszystkie lokalne stosunki szans poza tymi, które dotyczą komórek na głównej przekątnej są równe 1, tj:

$$\Theta_{ij}^{XY} = 1 \tag{3.40}$$

dla każdej pary kategorii $i$,$j$, takich, że $i \neq j$. Natomiast pozostałe stosunki szans, wyznaczone dla komórek na głównej przekątnej tj. $\Theta_{qq}^{XY}$ mogą być specyficzne. Hipotezę powyższą ilustruje tabela 3.41, można zauważyć, że interakcja zależy od tego

jakie kategorie dzielą kategorie $i$ oraz $j$. Ilustracja ta jest zgodna zgodna z następującą parametryzacją:

$$\pi_{ij}^{XY} = \begin{cases} d \cdot d_i^X \cdot d_j^Y \cdot \prod_{q=i}^{j-1} c_q, & \text{dla } i < j \\ \\ d \cdot d_i^X \cdot d_j^Y \cdot \prod_{q=j}^{i-1} c_q, & \text{dla } i > j \end{cases} \tag{3.41}$$

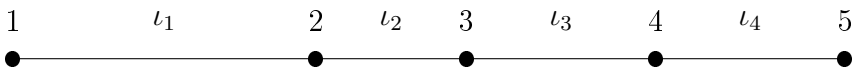
Tabela 3.41: Ilustracja modelu przekraczania barier CP — parametry interakcji

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | $\mathbf{c_1}$ | $\mathbf{c_1 c_2}$ | $c_1c_2c_3$ | $c_1c_2c_3c_4$ |
| $x_2$ | $c_1$ | **1** | $\mathbf{c_2}$ | $c_2c_3$ | $c_2c_3c_4$ |
| $x_3$ | $c_1c_2$ | $c_2$ | 1 | $c_3$ | $c_3c_4$ |
| $x_4$ | $c_1c_2c_3$ | $c_2c_3$ | $c_3$ | 1 | $c_4$ |
| $x_5$ | $c_1c_2c_3c_4$ | $c_2c_3c_4$ | $c_3c_4$ | $c_4$ | 1 |

Weźmy pod uwagę wielkość $\Theta_{12}^{XY}$, komórki definiujące ten stosunek szans zostały wyróżnione w tabeli 3.41. Załóżmy na chwilę, że tablica ta ilustruje ruchliwość edukacyjną a kolejnymi kategoriami wykształcenia są 1. podstawowe, 2. niepełne średnie, 3. średnie, 4.niepełne wyższe, 5. wyższe. Wielkość $\Theta_{12}^{XY}$ porównuje dwie grupy wyróżnione ze względu na wykształcenie ojca – *podstawowe* ($x_1$) lub *niepełne średnie* ($x_2$) – ze względu na proporcję liczby osób o wykształceniu *niepełnym średnim* ($y_2$) do liczby osób z wykształceniem *średnim* ($y_3$). Stosunek szans jest równy 1, co oznacza, że proporcja ta jest taka sama dla obydwu grup. Wynika to z tego, że bez względu na to czy bierze się pod uwagę osoby, których ojciec miał wykształcenie *podstawowe*, czy też *niepełne średnie* — osoby z wykształceniem średnim w porównaniu do osób z wykształceniem niepełnym średnim przekroczyły dodatkowo „barierę" *niepełne średnie–średnie* względem pozycji ich ojca, której odzwierciedleniem w tabeli 3.41 jest parametr $c_2$. Nie ma znaczenia, że osoby, których ojciec miał wykształcenie podstawowe musiały przekroczyć również „barierę" *podstawowe–niepełne średnie*, której odzwierciedleniem w tabeli 3.41 jest parametr $c_1$. Ponieważ barierę tę musiały przekroczyć zarówno osoby z wykształceniem niepełnym średnim jak i średnim, w modelu tym zakłada się, że nie ma to wpływu na rozpatrywaną proporcję między liczbą osób z tych dwóch grup.

Właśnie to założenie stanowi o istocie rozpatrywanego modelu. Na prawdopodobieństwo przekroczenia przez daną osobę bariery pomiędzy dwoma szczeblami wykształcenia - zgodnie z tym modelem - nie ma wpływu, czy przekracza ona dodatkowo

Rysunek 3.1: Odległości pomiędzy kolejnymi kategoriami zmiennej w modelu CP




inną barierę. Ujmując to inaczej bariery można zilustrować jako dystanse na jednowymiarowej przestrzeni. Ilustruje to rysunek 3.1. Bariera pomiędzy kategoriami *pierwszą* i *trzecią* jest sumą barier pomiędzy kategorią *pierwszą* i *drugą* oraz *drugą* i *trzecią* — oznaczonymi odpowiednio jako $\iota_1$ i $\iota_2$. To samo można powiedzieć o dowolnej parze kategorii. Dystanse $\iota_i$ można traktować jako parametry addytywnej wersji modelu przekraczania barier, tj. w stosunku do formuły 3.41 i tabeli 3.41 zachodzi $\iota_i = \log c_i$, tak więc parametr multiplikatywny opisujący barierę pomiędzy kategoriami *pierwszą* i *trzecią* jest iloczynem parametrów $c_1$ i $c_2$.

Oczywiście, założenie to można kwestionować, np. w pewnych sytuacjach rozsądne może być przypuszczenie, że dystans pomiędzy kategorią pierwszą i trzecią jest mniejszy aniżeli suma dwóch dystansów $\iota_1$ i $\iota_2$, można na przykład przypuszczać, że przekroczenie pierwszej bariery ułatwia przekroczenie następnej. Warto jednak mieć świadomość, że właśnie to założenie stanowi o istocie modelu i przekłada się na to, że odpowiednie lokalne stosunki szans są równe 1.

Warto jeszcze raz podkreślić, że jedynie stosunki szans $\Theta_{qq}^{XY}$ nie muszą w tym modelu być równe 1, gdyż zdają sprawę z trudności przekraczania barier pomiędzy kolejnymi kategoriami. Ponadto, warto zauważyć, że stosunki szans wyznaczone dla komórek znajdujących się na głównej przekątnej z wierszami i kolumnami oddalonymi o $k$–kategorii są iloczynem stosunków szans typu $\Theta_{qq}^{XY}$. Formalnie

$$\Theta_{i/(i+k);i/(i+k)}^{X\ \ Y} = \prod_{q=i}^{i+k-1} \Theta_{qq}^{XY} = \prod_{q=i}^{i+k-1} \left(\frac{1}{c_q}\right)^2 \tag{3.42}$$

dla każdego $i$ oraz dla każdego $k > 0$, przy spełnionym warunku $2 \leq i + k \leq r$.

Jak pokazuje warunek 3.40, model przekraczania barier jest szczególnym przypadkiem modelu symetrii. Wszystkie stosunki szans $\Theta_{ij}$, gdy $i \neq j$ są symetryczne, co więcej są one równe 1. Nie da się w ramach tego modelu uwzględnić przypuszczenia, że szanse na przekraczanie barier „w dół" są inne niż szanse na przekraczanie analogicznej bariery „w górę". Gdyby zmodyfikować parametry z tabeli 3.41 w ten

sposób, że parametry $c_1$, $c_2$, $c_3$ zastąpilibyśmy parametrami $c_1'$, $c_2'$, $c_3'$, to nie zmieniłoby to rozważanego modelu. Pozostałby on symetryczny, gdyż wszystkie stosunki szans nadal byłyby symetryczne.

Nie można natomiast porównać modelu CP z symetryczną wersją modelu zmiennego dystansu, tj. DS pod względem „prostoty" modelu. Z jednej strony model CP nie zakłada, że dystanse pomiędzy kolejnymi kategoriami zmiennej porządkowej są takie same. Z drugiej strony model CP zakłada, że dystanse te dają się przedstawić na jednym wymiarze. Model dystansu na odwrót: czyni pierwsze założenie, natomiast nie czyni drugiego tj. niekoniecznie zachodzi $s_3 = s_1 \cdot s_2$.

Zauważmy też, że zgodnie z warunkiem 3.40 liczba lokalnych stosunków szans równych 1, wynosi:

$$df = 2[(r-2) + (r-3) + \ldots + 1] = (r-2)(r-1).$$

co określa liczbę stopni swobody modelu CP. W stosunku do niezależności stochastycznej model ten posiada $r-1$ parametrów opisujących „bariery" pomiędzy kolejnymi kategoriami.

Zgodnie z metodą największej wiarygodności oszacowane rozkłady brzegowe — podobnie jak w większości modeli prezentowanych do tej pory — odzwierciedlają dane w próbie. Ponadto dla modelu CP, spełniony jest warunek:

$$\sum_{i}^{r}\sum_{j}^{r} \alpha_q \cdot \widehat{\pi}_{ij}^{XY} = \sum_{i}^{r}\sum_{j}^{r} \alpha_q \cdot p_{ij}^{XY} \tag{3.43}$$

dla każdego $1 \leq q \leq (r-1)$, gdzie $\alpha_q$ jest równe 1, dla takich kombinacji $x_i$, $y_j$ obydwu zmiennych, że spełniony jest jeden z dwóch warunków:

$$i \leq q < j, \text{ gdy } j > i,$$

bądź

$$j \leq q < i, \text{ gdy } j < i.$$

W pozostałych przypadkach $\alpha_q = 0$. Przykładowo $\alpha_2$=1, dla takich komórek powyżej przekątnej, że $i \leq 2$ oraz $j > 2$ jak również dla komórek poniżej przekątnej, gdy $j \leq 2$ oraz $i > 2$, czyli dla tabeli o wymiarach 5 na 5 są to komórki związane z prawdopodobieństwami $\pi_{13}^{XY}$, $\pi_{14}^{XY}$, $\pi_{15}^{XY}$, $\pi_{23}^{XY}$, $\pi_{24}^{XY}$, $\pi_{25}^{XY}$ oraz $\pi_{31}^{XY}$, $\pi_{41}^{XY}$, $\pi_{51}^{XY}$, $\pi_{32}^{XY}$, $\pi_{42}^{XY}$, $\pi_{52}^{XY}$. Warunek powyższy wskazuje, że suma oszacowanych prawdopodobieństw (liczebności) komórek opisujących przekroczenie pewnej bariery musi być zgodna dla rozkładu oczekiwanego i danych z próby, np. dla tabeli 3.42 zgodnej z modelem CP spełniony jest warunek dotyczący wszystkich komórek opisujący przejście pomiędzy

kategorią *z wielką trudnością* do kategorii *z trudnością*.

$$\begin{aligned} \widehat{\pi}_{12}^{XY} + \widehat{\pi}_{13}^{XY} + \widehat{\pi}_{14}^{XY} + \widehat{\pi}_{15}^{XY} + \widehat{\pi}_{21}^{XY} + \widehat{\pi}_{31}^{XY} + \widehat{\pi}_{41}^{XY} + \widehat{\pi}_{51}^{XY} &= \\ p_{12}^{XY} + p_{13}^{XY} + p_{14}^{XY} + p_{15}^{XY} + p_{21}^{XY} + p_{31}^{XY} + p_{41}^{XY} + p_{51}^{XY} &= 478,5. \end{aligned}$$

Dla powyższych komórek $\alpha_1 = 1$, podobny warunek możemy zdefiniować dla komórek dla których $\alpha_2 = 1$, $\alpha_3 = 1$, $\alpha_4 = 1$. Zauważmy ponadto, że skoro suma prawdopodobieństw dla komórek w pierwszym wierszu i pierwszej kolumnie z pominięciem prawdopodobieństwa $\widehat{\pi}_{11}^{XY}$ musi odzwierciedlać dane z próby, podobnie częstości z próby odzwierciedlają prawdopodobieństwa brzegowe $\widehat{\pi}_1^X$ oraz $\widehat{\pi}_1^Y$, to pośrednio wynika z tego, że $\widehat{\pi}_{11}^{XY} = p_{11}^{XY}$. Podobnie jest dla ostatniej komórki głównej przekątnej, tj. $\widehat{\pi}_{rr}^{XY} = p_{rr}^{XY}$.

Wyniki z tabeli 3.36 pokazują, że model CP jest słabo dopasowany w odniesieniu do obydwu analizowanych tablic. Tabela 3.42 — zamieszczona w celach ilustracyjnych — przedstawia rozkład oczekiwany zgodny z tym modelem dla danych panelowych. Odpowiednie parametry dla tego modelu wynoszą $c_1 = 0,53$, $c_2 = 0,60$, $c_3 = 0,49$, $c_4 = 0,36$. W addytywnej postaci parametry te wynoszą $\iota_1 = -0.64$, $\iota_2 = -0.51$, $\iota_3 = -0.70$, $\iota_4 = -1.01$. Ich wartości bezwzględne wskazują na trudność w przekraczaniu kolejnych barier, odzwierciedlających poprawę sytuacji materialnej. Powyższe wartości wskazują, że poprawa sytuacji z kategorii „z pewnym trudem" w 2000 roku na kategorię „raczej łatwo" w 2005 roku — parametr $c_3$ — nie jest tak trudne jak analogiczna poprawa sytuacji z kategorii „raczej łatwo" na kategorię „łatwo" — parametr $c_4$. Parametr $c_3$ jest bliższy wartości 1, a wyrażając to za pomocą addytywnych parametrów wielkość bezwzględna parametru $\iota_4$ jest większa niż parametru $\iota_3$.

Zgodnie z tym modelem, wszystkie stosunki szans w tej tabeli są równe 1, poza tymi, które są wyznaczone dla komórek znajdujących się na głównej przekątnej, np.

$$\Theta_{11}^{XY} = \frac{271,7 \cdot 171,2}{144,9 \cdot 89,6} = \frac{1}{c_1^2} = \frac{1}{0,53^2} = 3,58.$$

Z jednej strony wielkość powyższa uwzględnia proporcję liczby gospodarstw, które w zarówno w 2000 jak i w 2005 roku radziły sobie z „z wielką trudnością" do liczby gospodarstw, których sytuacja w tym okresie poprawiła się z kategorii „z wielką trudnością" na kategorię „z trudnością". Z drugiej strony uwzględnia proporcję liczby gospodarstw, w których sytuacja zmieniła się w odwrotną stronę do liczby gospodarstw zarówno w 2000 jak i w 2005 roku radziły sobie z „trudem". Pierwsza z tych proporcji jest ponad 3,5 razy większa niż druga. Analogiczny stosunek szans $\Theta_{22}^{XY}$ wynosi 2,76. Z modelu CP wynika, że stosunek szans porównujący pierwszą z trzecią kategorią obydwu zmiennych jest iloczynem powyższych lokalnych stosunków szans, tj.

$$\Theta_{1/3;1/3}^{X\ \ Y} = \Theta_{11}^{XY} \cdot \Theta_{22}^{XY} = 3,58 \cdot 2,76,$$

co jest zgodne z formułą 3.42. Jeśli chodzi o parametry przekraczania barier dla tablicy ruchliwości wynoszą one: $c_1 = 0,47$, $c_2 = 0,51$, $c_3 = 0,47$. Jak widać są one zbliżone do siebie.

Tabela 3.42: Rozkład oczekiwany zgodny z modelem CP dla danych z tabeli 3.6

| Czy przy aktualnym dochodzie netto Pana(i) gospodarstwo domowe wiąże koniec z końcem? | | | | | | |
|---|---|---|---|---|---|---|
| Odpowiedzi w 2000 roku ($X$) | Odpowiedzi w 2005 roku ($Y$) | | | | | Suma |
| | 1 | 2 | 3 | 4 | 5 | |
| 1. Z wielką trudnością | 271,7 | 144,9 | 123,1 | 36,1 | 6,4 | 582,2 |
| 2. Z trudnością | 89,6 | 171,2 | 145,5 | 42,6 | 7,6 | 456,5 |
| 3. Z pewną trudnością | 61,0 | 116,7 | 273,8 | 80,2 | 14,2 | 545,9 |
| 4. Raczej łatwo | 15,5 | 29,7 | 69,7 | 84,2 | 14,9 | 214,0 |
| 5. Łatwo | 1,9 | 3,5 | 8,3 | 10,0 | 13,5 | 37,2 |
| Suma | 439,7 | 466,0 | 620,4 | 253,1 | 56,6 | 1835,8 |

Przypomnijmy, że powyższe wyniki dotyczą modelu, którego dopasowane do danych okazało się słabe. Model ten jest często modyfikowany przez uwzględnienie specyfik głównej przekątnej (oznaczamy go wówczas QCP). Oczywiście, wówczas liczba niezależnych warunków dotyczących stosunków szans jest mniejsza. Dokładniej, w odniesieniu do lokalnych stosunków szans położonych na pseudo–przekątnych znajdujących się najbliżej głównej przekątnej, zakładamy jedynie, że są one symetryczne, tj. $\Theta^{XY}_{q(q+1)} = \Theta^{XY}_{(q+1)q}$ (porównaj tabelę 3.43). Oryginalna postać modelu (3.40)zakładała jedynie, że są równe 1, tak więc liczba warunków redukuje się z $2(r-2)$ do $(r-2)$. Liczba stopni zmniejsza się o $(r-2)$ i wynosi $df = (r-2)(r-2)$. Pokazuje to, że uwzględnienie specyfiki głównej przekątnej wymaga jedynie $(r-2)$ niezależnych parametrów. Można więc w odniesieniu do tabeli 3.41 założyć, że parametry $q_1$ oraz $q_5$ są równe 1[27]. Przypomnijmy, że komórki te przy szacowaniu rozkładu oczekiwanego zgodnego z modelem CP metodą największej wiarygodności, odzwierciedlają dane z próby, np. w tabeli 3.42 zgodnej z tym modelem, liczebności $\widehat{f}_{11}$, $\widehat{f}_{55}$ odzwierciedlają dokładnie liczebności próby.

Modele QCP są dobrze dopasowane do analizowanych danych. Uwzględnienie głównej przekątnej w obydwu przypadkach poprawia istotnie dopasowanie statystyczne modelu, tj. wyniki testów warunkowych są następujące: $G^2 = 29,2$, $df = 2$, $(p < 0,0001)$ dla tablicy ruchliwości społecznej oraz $G^2 = 43,0$, $df = 3$ $(p < 0,0001)$

[27] Alternatywnie można założyć, np. że $q_1 = q_2$ oraz $q_4 = q_5$.

Tabela 3.43: Ilustracja modelu przekraczania barier QCP — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $q_1$ | $c_1$ | $c_1c_2$ | $c_1c_2c_3$ | $c_1c_2c_3c_4$ |
| $x_2$ | $c_1$ | $q_2$ | $c_2$ | $c_2c_3$ | $c_2c_3c_4$ |
| $x_3$ | $c_1c_2$ | $c_2$ | $q_3$ | $c_3$ | $c_3c_4$ |
| $x_4$ | $c_1c_2c_3$ | $c_2c_3$ | $c_3$ | $q_4$ | $c_4$ |
| $x_5$ | $c_1c_2c_3c_4$ | $c_2c_3c_4$ | $c_3c_4$ | $c_4$ | $q_4$ |

dla danych panelowych. W pierwszym przypadku parametry modelu wynoszą $c_1 = 0,42$, $c_2 = 0,40$, $c_3 = 0,44$, a odpowiednie parametry przekątnej $q_2 = 0,58$, $q_3 = 0,46$, natomiast dla danych panelowych $c_1 = 0,50$, $c_2 = 0,47$, $c_3 = 0,34$, $c_4 = 0,37$, a odpowiednie parametry przekątnej $q_2 = 0,52$, $q_3 = 0,61$, $q_4 = 0,45$.

Zwraca uwagę, że parametry głównej przekątnej są mniejsze od 1, jednak nie wynika z tego, że współwystępowanie tych samych kategorii obydwu zmiennych, np. dziedziczenie wykształcenia średniego jest relatywnie mało prawdopodobne. Tak jak w przypadku modelu quasi–niezależności, bądź innych modeli uwzględniających specyfikę głównej przekątnej — model QCP można sobie wyobrazić jako wyszczególnienie osób niemobilnych ($Z = 1$) i dwie podzbiorowości osób mobilnych, dla których związek między $X$ oraz $Y$ opisuje model przekraczania barier, przy czy dla jednej z nich dopuszczamy możliwość dziedziczenia pozycji ($Z = 2$) a dla kolejnej wykluczamy taką sytuację ($Z = 3$). Wartości parametrów głównej przekątnej mniejsze od 1 nie oznaczają, że istnieje tendencja odwrotna do dziedziczenia pozycji. Wartości te ustalone są względem hipotetycznej sytuacji modelu CP a nie modelu niezależności. Wynik ten oznacza jedynie, że dla podzbiorowości $Z = 2$ odsetek osób ulokowanych na głównej przekątnej jest zawyżony w stosunku do całej zbiorowości.

### Model ustalonego dystansu FD

Model ustalonego dystansu — oznaczany jako FD — został sformułowany przez Habermana (1974*b*, 1979). Jest on zagnieżdżony zarówno w symetrycznej wersji modelu dystansu jak również w modelu przekraczaniu barier. Definiuje go połączenie warunków 3.35 oraz 3.40.

$$\Theta_{ij}^{XY} = 1 \quad \text{oraz} \quad \Theta_{qq}^{XY} = const, \tag{3.44}$$

dla każdej pary kategorii $i$, $j$, takich, że $i \neq j$ oraz $1 \leq i \leq (r-1)$, $1 \leq j \leq (r-1)$, oraz dla każdego $q$, takiego, że $1 \leq q \leq r$. Parametryzację tego modelu można przedstawić

następująco:

$$\pi_{ij}^{XY} = d \cdot d_i^X \cdot d_j^Y \cdot s^k \qquad \text{gdzie } k = |j - i|. \tag{3.45}$$

Ilustrację modelu FD stanowi tablica 3.44. Zgodnie z tym modelem interakcja zależy od tego, o ile dana komórka jest oddalona od głównej przekątnej. Parametr interakcji $s_k$ jest „proporcjonalny" do liczby kategorii jakie dzielą obydwie zmienne. W przypadku ruchliwości wskazuje to, o ile komórek nastąpił „awans" bądź „degradacja" wykształcenia syna w stosunku do wykształcenia ojca.

Model ten łączy więc założenia dwóch poprzednio omawianych modeli, dotyczące traktowania różnicy pomiędzy wartościami obydwu analizowanych zmiennych. Z jednej strony zakłada się w nim, że dystanse pomiędzy kolejnymi kategoriami zmiennej porządkowej są takie same, z drugiej strony, że dystanse te dają się przedstawić na jednym wymiarze. Zauważyć należy, że w stosunku do modelu dystansu $s_k = s^k$, natomiast w stosunku do modelu przekraczania barier $\prod_{q=i}^{i+k-1} c_q = s^k$, czyli na przykład $c_1 \cdot c_2 = s^2$.

Tabela 3.44: Ilustracja modelu ustalonego dystansu FD — parametry interakcji

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | $s$ | $s^2$ | $s^3$ | $s^4$ |
| $x_2$ | $s$ | 1 | $s$ | $s^2$ | $s^3$ |
| $x_3$ | $s^2$ | $s$ | 1 | $s$ | $s^2$ |
| $x_4$ | $s^3$ | $s^2$ | $s$ | 1 | $s$ |
| $x_5$ | $s^4$ | $s^3$ | $s^2$ | $s$ | 1 |

Liczba stopni swobody tego modelu wynosi $df = (r-1)^2 - 1 = r(r-2)$, co jest zgodne z liczbą warunków nakładanych na lokalne stosunki szans. Zakłada się, że $(r-2)(r-1)$ stosunków szans jest równych 1, a stosunki szans na głównej przekątnej są sobie równe, co daje $(r-2)$ dodatkowych warunków. Pokazuje to, że model FD w stosunku do niezależności stochastycznej posiada tylko jeden parametr więcej, mianowicie $s$.

Model ten — podobnie jak poprzednie — modyfikuje się często uwzględniając parametry różnicujące komórki na głównej przekątnej (oznaczamy go wówczas QFD), tak jak w tabeli 3.45. Uwzględnienie parametrów głównej przekątnej wymaga $r$ parametrów, inaczej niż w stosunku do modelu D, DS, czy modelu CP. Liczba stopni swobody dla tego modelu wynosi więc:

$$df = r^2 - 3r.$$

Tabela 3.45: Ilustracja modelu ustalonego dystansu QFD — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $q_1$ | $s$ | $s^2$ | $s^3$ | $s^4$ |
| $x_2$ | $s$ | $q_2$ | $s$ | $s^2$ | $s^3$ |
| $x_3$ | $s^2$ | $s$ | $q_3$ | $s$ | $s^2$ |
| $x_4$ | $s^3$ | $s^2$ | $s$ | $q_4$ | $s$ |
| $x_5$ | $s^4$ | $s^3$ | $s^2$ | $s$ | $q_5$ |

Sformułowanie tego modelu w języku lokalnych stosunków szans, jak również bardziej precyzyjne uzasadnienie liczby stopni swobody dla tego modelu można znaleźć w Aneksie.

To, że model QFD wymaga $r$ parametrów przekątnej a nie $(r-1)$ jak w innych modelach dystansu wynika to z tego, że model ten uwzględnia specyfikę głównej przekątnej w szczególny sposób. O ile parametr $s < 1$, to model FD rzeczywiście głosi, że komórki na głównej przekątnej występują relatywnie częściej niż komórki na innych pseudo–przekątnych, przy kontroli rozkładów brzegowych. Zakładamy jednak, że parametry interakcji zmieniają się na kolejnych pseudo–przekątnych „proporcjonalnie", co może nie być realistycznym założeniem. Model FD nie czyni takiego założenia. Jeśli chcemy uwzględnić faktyczną tendencję do dziedziczenia w modelu QFD wymaga to więc dodatkowego parametru.

O ile model FD nie jest dobrze dopasowany do analizowanych danych, to wyniki dla modelu QFD, są w obydwu przypadkach zadowalające. Parametr $s = 0,4$ dla tablicy ruchliwości, natomiast kolejne parametry przekątnej wynoszą $q_1 = 0,92$, $q_2 = 0,58$ $q_3 = 0,45$, $q_4 = 0,83$. Podobnie — jak w poprzednio omawianych modelach — wielkości mniejsze od 1 nie wskazują na wysoką mobilność i relatywnie małe prawdopodobieństwo dziedziczenia wykształcenia po ojcu, a jedynie na to, że w modelu FD ta tendencja była relatywnie silnie zarysowana. Dla danych panelowych wyniki są następujące $s = 0,43$, natomiast kolejne parametry przekątnej wynoszą $q_1 = 0,77$, $q_2 = 0,49$ $q_3 = 0,59$, $q_4 = 0,68$, $q_5 = 1,50$.

Jak zostało podkreślone na początku model FD jest zagnieżdżony w modelach CP i modelu DS, analogicznie model QFD jest szczególnym przypadkiem modeli QCP i QDS. Porównanie tych modeli za pomocą testów warunkowych pozwala zweryfikować założenie dotyczące możliwości określania wielkości stosunku szans w zależności od różnicy w kategoriach zmiennej wierszowej i kolumnowej (porównanie modeli CP i FD, bądź QCP i QFD) bądź założenie o addytywności barier (porównanie modeli

DS i FD, bądź QDS i QFD). Pierwsza z hipotez w odniesieniu do tablicy ruchliwości nie może być odrzucona, jeśli porównujemy model CP i FD, tj. $G^2 = 0,74$, $df = 2$ ($p = 0,69$), choć w tym przypadku test jest kontrowersyjny ze względu na słabe dopasowanie modelu CP. Dopasowanie modeli QCP i QFD jest identyczne, ale nie jest to wynik przypadkowy — modele te dla tablicy o wymiarach 4 x 4 są tożsame, co zostało pokazane w Aneksie.

Jeśli chodzi o dane panelowe to porównanie modeli CP i FD daje następujące wyniki: $G^2 = 9,4$, $df = 3$ ($p = 0,024$), przy czym model CP jest słabo dopasowany do danych co czyni test wątpliwym. Dla tych danych można porównać modele QCP i QFD: $G^2 = 4,51$, $df = 1$ ($p = 0,03$). Jak widać, na poziomie istotności 0,01 hipoteza o możliwości określenia wielkości stosunku szans w zależności od liczby kategorii dzielących zmienną wierszową i kolumnową nie może być odrzucona.

Jeśli chodzi o drugą hipotezę — addytywności barier — to w odniesieniu do tablic ruchliwości inne są konkluzje, jeśli porównujemy modele DS i FD i modele QDS i QFD: w pierwszym przypadku $G^2 = 25,0$, $df = 2$ ($p < 0,0001$), a w drugim przypadku $G^2 = 0,15$, $df = 1$ ($p = 0,69$). Drugi z testów wydaje się bardziej uprawniony, gdyż przyjęcie, że model DS jest realistyczny jest kontrowersyjne jeśli weźmie się pod uwagę dopasowanie tego modelu. Ponadto warto zauważyć, że znaczna redukcja statystyki $G^2$ wynika w pewnej mierze, z tego, że model DS uwzględnia specyfikę głównej przekątnej, inaczej niż w modelu FD. Do podobnych konkluzji prowadzi przeprowadzenie analogicznych testów dla danych panelowych: $G^2 = 41,4$, $df = 3$ ($p < 0,0001$) dla modeli DS i FD oraz $G^2 = 1,2$, $df = 2$ ($p = 0,54$) dla modeli QDS i QFD.

### Inne modele dystansu

W literaturze pojawiło się wiele modyfikacji modeli dystansu. Dla przykładu Smits i inni (1998) w swoich badaniach dotyczących homogamii małżeńskiej przedstawiają model, w którym stosunki szans — podobnie jak w modelu zmiennego dystansu i ustalonego dystansu — zależą od tego, której pseudo–przekątnej dotyczą. Przypomnijmy, że w modelu ustalonego dystansu są one równe 1 (poza główną przekątną), a odpowiednie parametry interakcji wynoszą $s^k$, w tym sensie parametry interakcji dla kolejnych pseudo–przekątnych maleją lub rosną „proporcjonalnie". Mówiąc dokładniej parametry w wersji addytywnej maleją liniowo względem liczby kategorii, które dzielą obydwie zmienne: wynoszą one $\eta k$, gdzie $\eta = \ln(s)$. W zaproponowanej modyfikacji parametry te wynoszą $s^{k^{1,5}}$, tak więc w wersji addytywnej parametry maleją lub rosną nieliniowo dla kolejnych pseudo–przekątnych, wynoszą one $\eta k^{1,5}$. Mówiąc w pewnym uproszczeniu, w modelu tym szanse na awans bądź degradację nie zmieniają się tak

„proporcjonalnie", jak w modelu FD, względem liczby kategorii, których ten awans dotyczy.

Tabela 3.46: Ilustracja modelu ustalonego dystansu $FD^{1,5}$ — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | $s$ | $s^{2^{1,5}}$ | $s^{3^{1,5}}$ | $s^{4^{1,5}}$ |
| $x_2$ | $s$ | 1 | $s$ | $s^{2^{1,5}}$ | $s^{3^{1,5}}$ |
| $x_3$ | $s^{2^{1,5}}$ | $s$ | 1 | $s$ | $s^{2^{1,5}}$ |
| $x_4$ | $s^{3^{1,5}}$ | $s^{2^{1,5}}$ | $s$ | 1 | $s$ |
| $x_5$ | $s^{4^{1,5}}$ | $s^{3^{1,5}}$ | $s^{2^{1,5}}$ | $s$ | 1 |

Ilustrację tego modelu — będzie on oznaczany jako $\mathrm{FD}^{1,5}$ — stanowi tabela 3.46. Jak widać model ten — podobnie jak model ustalonego dystansu — wykorzystuje tylko jeden parametr więcej aniżeli model niezależności stochastycznej. Zgodnie z tym modelem lokalne stosunki szans wynoszą:

$$\Theta^{XY}_{i(i+k)} = s^{2k^{1,5}-(k-1)^{1,5}-(k+1)^{1,5}} \text{ dla } k \neq 0 \tag{3.46}$$

$$\Theta^{XY}_{ii} = 1/s^2 \tag{3.47}$$

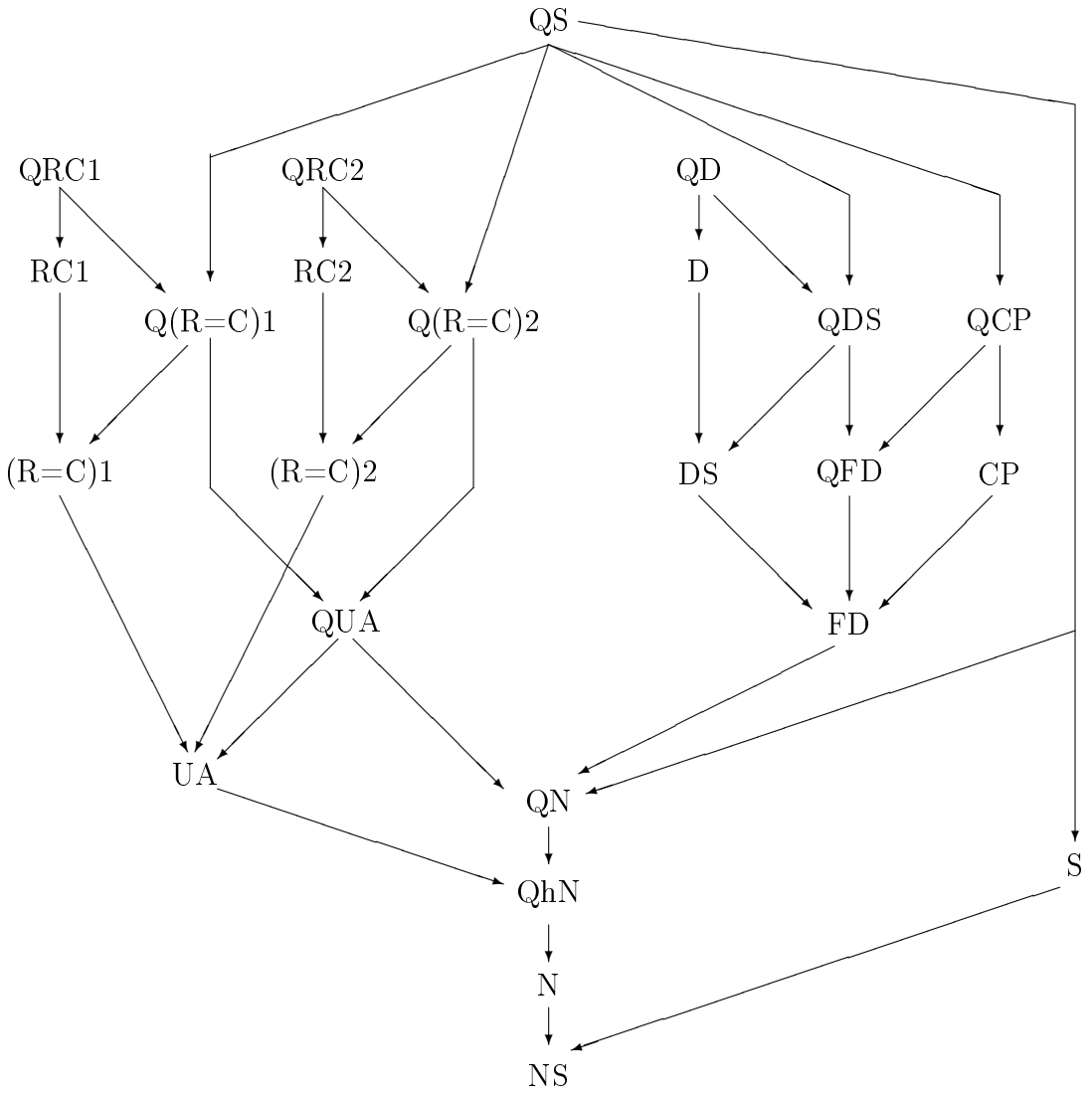
przy czym $i = 1, \ldots r-1$, $k = -(r-2), \ldots r-2$ oraz $i + k \leq r-1$. W porównaniu do modelu zmiennego dystansu stosunki te nie są specyficzne dla każdej pseudo–przekątnej, omawiany model jest więc od niego prostszy. W podobny sposób można formułować inne modele, na przykład podnosząc wielkość $k$ do innej potęgi.

W tabeli 3.36 przedstawione zostały wyniki dopasowania dla tego modelu $\mathrm{FD}^{1,5}$, również w wersji zmodyfikowanej przez dodanie parametrów głównej przekątnej $\mathrm{QFD}^{1,5}$. Jak widać, model ten nawet w wersji prostszej $\mathrm{FD}^{1,5}$ stanowi akceptowalny opis zarówno tablicy ruchliwości edukacyjnej jak też danych panelowych.

### 3.2.5 Podsumowanie omówionych modeli

Warto podsumować modele omówione do tej pory. Zostały one przedstawione na rysunku 3.2. Strzałki wskazują na relacje, jakie zachodzą pomiędzy poszczególnymi modelami, kierunek wskazuje na przejście pomiędzy modelem bardziej złożonym a prostszym, np. dodając do modelu QS założenie o identyczności rozkładów brzegowych, można sformułować model S. Zagnieżdżenie modeli jest relacją przechodnią,

Rysunek 3.2: Relacje pomiędzy wybranymi modelami przedstawionymi w rozdziale trzecim



stąd też jeśli od modelu A możemy przejść za pomocą strzałek— zgodnie z ich kierunkiem — do modelu B, oznacza to, że model B jest zagnieżdżony w modelu A. Na przykład model N jest zagnieżdżony[28] w modelu QS, ale model QDS nie jest zagnieżdżony ony w modelu QCP.

Jak widać prawie wszystkie modele — poza RC1, RC2, QRC1, QRC2, D, QD — są zagnieżdżone w modelu QS, tak więc stosunki szans są w nich symetryczne. Widać również, że modele jednakowej interakcji i wierszowo–kolumnowe wychodzą z zupełnie innych założeń niż modele dystansu, podobnie inne założenia występują w modelach przekraczania barier. Mówiąc inaczej, te grupy modeli opisują inny rodzaj związku między zmiennymi. Model ustalonego dystansu FD jest szczególnym przypadkiem modeli dystansu i przekraczania barier. Na rysunku uwzględnione zostały również modyfikacje różnych modeli pozwalające na uwzględnienie komórek głównej przekątnej. Model quasi–niezależności jest szczególnym przypadkiem tych modeli, np. modelu QFD.

W rozdziale trzecim uwzględniane były również inne modele, które nie zostały uwzględnione na rysunku, aby nie utrudniać jego percepcji. Na przykład nie zostały uwzględnione modele wierszowe R, QR, C, CR, $FD^{1,5}$. Przykładowo, model QR1 można umieścić pomiędzy modelem QRC1 i modelem QUA.

## 3.3 Modelowanie asymetrii

Na początku tego rozdziału poruszony został problem asymetrii, tj. częstszego występowania komórek powyżej lub poniżej głównej przekątnej. Kwestia ta jest szczególnie istotna w odniesieniu do zmiennych porządkowych, gdyż jedynie wówczas podział na komórki powyżej i poniżej przekątnej jest merytorycznie uzasadniony. Przykładowo, istotna może być odpowiedź na pytanie, czy częściej zdarzają się respondenci, którzy mają wyższe wykształcenie niż ich ojcowie, aniżeli tacy, którzy są od swoich ojców wykształceni gorzej. Podobnie w odniesieniu do danych panelowych sensowne jest porównanie odsetka osób, których sytuacja materialna polepszyła się w stosunku do odsetka osób, których sytuacja uległa pogorszeniu.

Przypomnijmy, że w literaturze dotyczącej struktury społecznej rozróżnia się na ogół ruchliwość absolutną (obejmująca również zmiany strukturalne) od ruchliwości względnej. W podobny sposób można mówić o absolutnej i względnej asymetrii. Odsetki przytoczone w powyższym akapicie zdają sprawę z pierwszej z nich, natomiast nie odzwierciedlają one asymetrii względnej, gdyż na ich wielkość istotny wpływ mają

[28]Przypomnijmy, że model prostszy jest zagnieżdżony w modelu bardziej złożonym, tj. jeśli model A jest zagnieżdżony w modelu B, ten ostatni posiada pewne dodatkowe założenia.

różnice w rozkładach brzegowych. Modele prezentowane w tej części dotyczyć będą względnej asymetrii. Dlatego też punktem wyjścia będzie raczej model quasi-symetrii aniżeli model symetrii, gdyż pierwszy z nich w odróżnieniu od drugiego nie zakłada nic na temat rozkładów brzegowych. Warunek quasi-symetrii najczęściej definiuje się w literaturze na dwa sposoby. Jeden z nich, wykorzystywany dotychczas w tej pracy, odwołuje się do stosunków szans, które — zgodnie z modelem QS — są symetryczne względem głównej przekątnej, tj.

$$\frac{\Theta_{ij}^{XY}}{\Theta_{ji}^{XY}} = 1 \tag{3.48}$$

W innym ujęciu — wprowadzonym przez Caussinusa (1965) — quasi-symetrię definiuje się za pomocą innego warunku:

$$\Phi_{ijk} = \frac{F_{ij}^{XY} \cdot F_{jk}^{XY} \cdot F_{ki}^{XY}}{F_{kj}^{XY} \cdot F_{ji}^{XY} \cdot F_{ik}^{XY}} = \frac{\pi_{(i)j}^{(X)Y} \cdot \pi_{(j)k}^{(X)Y} \cdot \pi_{(k)i}^{(X)Y}}{\pi_{(k)j}^{(X)Y} \cdot \pi_{(j)i}^{(X)Y} \cdot \pi_{(i)k}^{(X)Y}} = 1 \tag{3.49}$$

dla każdej trójki wartości $i, j, k$, takich, że $i < j < k$. Warto skupić się na zapisie wyrażonym w prawdopodobieństwach warunkowych. Przypomnijmy, że oznaczenie $\pi_{(i)j}^{(X)Y}$, oznacza prawdopodobieństwo, że zmienna $Y$ będzie miała wartość $j$, gdy zmienna $X$ ma wartość $i$. Przykładowo, dla analizowanych wcześniej danych panelowych wielkość $\pi_{(2)3}^{(X)Y}$ oznacza, prawdopodobieństwo, że respondent, który 2000 roku wskazał na odpowiedź „z trudnością", w roku 2005 odpowie „z pewną trudnością". Analogicznie dla danych ruchliwości edukacyjnej wielkość ta oznacza prawdopodobieństwo, że respondent, którego ojciec ma wykształcenie „niepełne średnie", sam uzyska wykształcenie „średnie". Wielkość w liczniku wskazuje na prawdopodobieństwo „cyklu" $i-j-k-i$. Zgodnie z założonym powyżej uporządkowaniem, na cykl ten składają się dwa przejścia „w górę", tj. $i-j$ oraz $j-k$, jak również jedno przejście „w dół" $k-i$. Cykl w mianowniku $i-k-j-i$ na odwrót: zawiera dwa przejścia „w dół" ($j-i$ oraz $k-j$) oraz jedno przejście „w górę" ($i-k$). Warunek 3.49 wskazuje, że obydwa cykle zdarzają się z takim samym prawdopodobieństwem.

Ilustrację wielkości $\Phi_{ijk}$ stanowi tablica 3.47, dla $i = 2$, $j = 3$, $k = 5$. Liczebności komórek w liczniku zostały oznaczone jako „$x$", natomiast liczebności komórek w mianowniku oznaczono symbolami „$o$". Zauważmy, że w drugim, trzecim i piątym wierszu mamy po jednej komórce oznaczonej „$x$" i po jednej komórce oznaczonej „$o$". Podobnie rzecz się ma z kolumnami dla interesujących nas kategorii. W konsekwencji, jeśli przemnożymy dowolny wiersz lub dowolną kolumnę przez jakąkolwiek stałą nie wpłynie to na wielkość $\Phi_{ijk}$. Oznacza to, że nie zależy ona — podobnie jak wielkości stosunków szans — od częstości brzegowych. Jak widać, różnice strukturalne, choć mogą mieć wpływ na poszczególne liczebności występujące w wyrażeniu 3.49, nie

Tabela 3.47: Ilustracja wielkości $\Phi_{235}$

| $X \backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | | | | | |
| $x_2$ | | | $x_{23}$ | | $o_{25}$ |
| $x_3$ | | $o_{32}$ | | | $x_{35}$ |
| $x_4$ | | | | | |
| $x_5$ | | $x_{52}$ | $o_{53}$ | | |

mają wpływu na wielkość $\Phi_{ijk}$. Gdyby była zachowana quasi-symetria, poszczególne parametry interakcji związane z komórkami położonymi symetrycznie wokół przekątnej — przykładowo komórkami oznaczonymi $x_{23}$ oraz $o_{32}$ — byłyby sobie równe i w związku z tym wielkość $\Phi_{ijk}$ wynosiłaby 1. Skoro nie jest równa 1 wskazuje to na asymetrię nie wynikającą z różnic strukturalnych[29].

Większość z modeli prezentowanych do tej pory stanowiło szczególny przypadek modelu quasi-symetrii, zakładało się w nich, że symetryczny jest związek a nie poszczególne prawdopodobieństwa. Jedynie kilka nie było modelami quasi–symetrycznymi, np. model wierszowy, heterogeniczne modele wierszowo–kolumnowe. W dalszej części nie tyle wymienione zostaną kolejne modele tego typu, co zaprezentowane zostaną ogólne możliwości modyfikacji modeli quasi–symetrii. Będą się one odwoływać bądź do ilorazu stosunków szans tj. wielkości z formuły 3.48 bądź porównania prawdopodobieństw wystąpienia dwóch cykli tj. wielkości $\Phi_{ijk}$ z formuły 3.49. Modyfikacje te będą polegały na tym, że wielkości powyższe nie są równe 1 a modelowane są za pomocą jednego lub większej liczby parametrów. Pierwszy typ modyfikacji prowadzi do sformułowania tzw. *asymetrycznych modeli związku* (*skew–symmetric association models*), natomiast hipotezy formułowane za pomocą wielkości $\Phi_{ijk}$ będziemy nazywać *asymetrycznymi modelami poziomu* (*skew–symmetric level models*).

Powyższy podział został zaproponowany przez Kazuo Yamaguchiego (1990), który szczegółowo omawia zagadnienie asymetrii. W tej pracy skoncentruję się na wybranych zaprezentowanych przez niego asymetrycznych modyfikacjach pierwszego i drugiego typu. Warto zaznaczyć, że uwzględnienie asymetrii niekoniecznie musi dotyczyć modyfikacji quasi–symetrii w najogólniejszej postaci. Modyfikowane w ten sposób mogą być rozmaite modele, które są szczególnymi przypadkami quasi–symetrii, np.

[29]Pomocne w zrozumieniu tych zagadnień może być rozróżnienie różnych typów ruchliwości (Lissowski 1991). Szczegółowe omówienie tych zagadnień przekracza jednak ramy tej pracy.

model jednakowej interakcji, symetryczny model wierszowo-kolumnowy, model symetrycznego dystansu, model przekraczania barier. Prezentując zagadnienie asymetrii, podane zostaną przykłady, pokazujące jak można modyfikować wybrane z powyższych modeli uwzględniając szczególne typy asymetrii. Przedstawiony zostanie przykład empiryczny pokazujący, że uwzględnienie asymetrii może poprawić dopasowanie do danych w sposób istotny statystycznie. Omówiona zostanie również interpretacja parametrów opisujących asymetrię.

### Asymetryczne modele związku

W modelach quasi–symetrii iloczyn dwóch stosunków szans zdefiniowanych dla komórek położonych symetrycznie względem głównej przekątnej, tj. $\Theta_{ij}^{XY}$ oraz $\Theta_{ji}^{XY}$ jest równy 1. W zaproponowanych przez Yamaquchiego modelach asymetrycznego związku iloraz ten jest definiowany przez parametr lub parametry asymetrii. W najprostszym modelu wykorzystuje się tylko jeden parametr, mianowicie:

$$\frac{\Theta_{ij}^{XY}}{\Theta_{ji}^{XY}} = a^2 \text{ dla } i < j \tag{3.50}$$

Ten typu asymetrycznej modyfikacji quasi–symetrii będzie nazywany modelem jednakowej asymetrycznej interakcji i oznaczany jako $U\_SK$. Zgodnie z tym modelem stosunki szans można przedstawić jako:

$$\Theta_{ij}^{XY} = \begin{cases} S_{ij}^{XY} \cdot a, & \text{dla } i < j \\ \\ S_{ij}^{XY} \cdot \frac{1}{a}, & \text{dla } i > j \end{cases} \tag{3.51}$$

gdzie $S_{ij}^{XY}$ wskazuje na komponent symetryczny, natomiast parametr $a$ na asymetryczny składnik lokalnego stosunku szans. Ilustrację tego modelu przedstawia tabela 3.48. Jest ona zgodna z następującą parametryzacją:

$$\pi_{ij}^{XY} = d \cdot d_i^X \cdot d_j^Y \cdot s_{ij}^{XY} \cdot a_{ij}^{XY} \tag{3.52}$$

przy czym kategoriami odniesienia dla obydwu zmiennych są wartości $x_1, y_1$. Wielkość $s_{ij}^{XY}$, jest symetrycznym parametrem interakcji tj. zachodzi $s_{ij}^{XY} = s_{ji}^{XY}$, a parametr asymetrii $a_{ij}^{XY}$ można zdefiniować jako:

$$a_{ij}^{XY} = \begin{cases} a^{(i-1)(j-i)}, & \text{dla } i < j \\ \\ \frac{1}{a^{(j-1)(i-j)}}, & \text{dla } i > j \end{cases} \tag{3.53}$$

Ponieważ ten typ asymetrii wymaga tylko jednego dodatkowego parametru, model ten ma $(r-1)(r-2)/2 - 1$ stopni swobody. Jednakowa asymetryczna interakcja w

Tabela 3.48: Ilustracja modelu quasi–symetrii z jednakową asymetryczną interakcją — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $s_{22}$ | $s_{23} \cdot a$ | $s_{24} \cdot a^2$ | $s_{25} \cdot a^3$ |
| $x_3$ | 1 | $s_{23} \cdot \frac{1}{a}$ | $s_{33}$ | $s_{34} \cdot a^2$ | $s_{35} \cdot a^4$ |
| $x_4$ | 1 | $s_{24} \cdot \frac{1}{a^2}$ | $s_{34} \cdot \frac{1}{a^2}$ | $s_{44}$ | $s_{45}a^3$ |
| $x_5$ | 1 | $s_{25} \cdot \frac{1}{a^3}$ | $s_{35} \cdot \frac{1}{a^4}$ | $s_{45} \cdot \frac{1}{a^3}$ | $s_{55}$ |

powyższym przypadku jest modyfikacją ogólnej postaci quasi–symetrii, ale - tak jak zostało zasygnalizowane wcześniej — można ją również odnieść do różnych prostszych hipotez zakładających warunek 3.48. Przykładowo, odnosząc ją do modelu jednakowej interakcji warunek dotyczący lokalnych stosunków szans można zdefiniować jako:

$$\Theta_{ij}^{XY} = \begin{cases} \delta \cdot a, & \text{dla } i < j \\ \\ \delta \cdot \frac{1}{a}, & \text{dla } i > j \end{cases} \tag{3.54}$$

Model taki oznaczać będziemy jako $UA\_U\_SK$. Ilustrację tego modelu stanowi tabela 3.49. Warto podkreślić, że interakcję w tym modelu opisują jedynie dwa parametry: symetryczny komponent $\delta$ oraz asymetryczny składnik $a$.

Tabela 3.49: Ilustracja modelu $UA\_U\_SK$ — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\delta$ | $\delta^2 \cdot a$ | $\delta^3 \cdot a^2$ | $\delta^4 \cdot a^3$ |
| $x_3$ | 1 | $\delta^2 \cdot \frac{1}{a}$ | $\delta^4$ | $\delta^6 \cdot a^2$ | $\delta^8 \cdot a^4$ |
| $x_4$ | 1 | $\delta^3 \cdot \frac{1}{a^2}$ | $\delta^6 \cdot \frac{1}{a^2}$ | $\delta^9$ | $\delta^{12} \cdot a^3$ |
| $x_5$ | 1 | $\delta^4 \cdot \frac{1}{a^3}$ | $\delta^8 \cdot \frac{1}{a^4}$ | $\delta^{12} \cdot \frac{1}{a^3}$ | $\delta^{16}$ |

Czy uwzględnienie asymetrii wydaje się użyteczne w stosunku do danych analizowanych do tej pory? Jeśli chodzi o dane panelowe opisujące ocenę sytuacji materialnej (tabela 3.6), to uwzględnienie asymetrii nie wydaje się konieczne. Wyniki z tabeli 3.11 pokazują, że quasi–symetria stanowi niemal idealny opis tych danych. Mniej jednoznaczne są wyniki dla danych ruchliwości edukacyjnej (tabela 3.5), niemniej na poziomie 0,01 hipoteza o quasi–symetrii nie może być odrzucona. Dla celów

ilustracyjnych w tabeli 3.50 przedstawiamy dodatkowo dane panelowe pochodzące z badania „Diagnoza Społeczna". Porównywane są odpowiedzi respondentów oceniających w 2003 i 2005 roku, jaki wpływ na ich życie miały zmiany, jakie zaszły w Polsce po roku 1989[30].

Tabela 3.50: Wpływ zmian w Polsce po 1989 roku na życie respondenta — porównanie odpowiedzi z 2003 i 2005 roku[a]

| | Odpowiedzi z 2005 roku | | | | | |
|---|---|---|---|---|---|---|
| Odpowiedzi z 2003 roku | 1 | 2 | 3 | 4 | 5 | Suma |
| 1. Bardzo niekorzystny | 199,7 | 300,3 | 148,6 | 43,2 | 4,5 | 696,3 |
| 2. Raczej niekorzystny | 181,2 | 759,1 | 495,5 | 207,8 | 17,5 | 1661,1 |
| 3. Brak wpływu | 132,1 | 491,0 | 1080,9 | 280,4 | 21,7 | 2006,1 |
| 4. Raczej korzystny | 28,8 | 106,9 | 149,7 | 259,7 | 48,9 | 594,0 |
| 5. Bardzo korzystny | 2,6 | 14,7 | 10,1 | 24,1 | 27,2 | 78,7 |
| Suma | 544,4 | 1672,0 | 1884,8 | 815,2 | 119,8 | 5036,2 |

[a]Źródło: Diagnoza społeczna, 2003–2005.

Dla tej tabeli model zakładający quasi–symetrię nie może być zaakceptowany na poziomie istotności $\alpha = 0,01$. Dla tego modelu statystyki dopasowania wynoszą odpowiednio: $\chi^2 = 18,3$ $(p = 0,0054)$ oraz $G^2 = 17,9$ $(p = 0,0064)$, przy 6 stopniach swobody. Indeks rozbieżności wynosi $\Delta = 1,62$. Uwzględnienie dodatkowo asymetrycznego związku w modelu quasi–symetrii czyni model akceptowalnym. Model będący połączeniem quasi–symetrii i jednakowej asymetrycznej interakcji jest akceptowalny: $\chi^2 = 6,11$ $(p = 0,2954)$ oraz $G^2 = 6,12$ $(p = 0,2943)$, $df = 5$. Indeks rozbieżności wynosi $\Delta = 0,09$. Poprawa dopasowania związana z uwzględnieniem jednego dodatkowego parametru asymetrii jest istotna statystycznie, co pokazują wyniki testu warunkowego: $G^2 = 18,3 - 6,1 = 12,2$ $(p = 0,0004)$, $df = 1$.

[30]Do konstrukcji zmiennej wykorzystane zostały dwa pytania. Pierwsze brzmiało: *Czy zmiany, jakie zaszły w Polsce od 1989 r. miały wpływ na Pana życie*? Respondenci, którzy odpowiedzieli „Tak" zostali dodatkowo zapytani: *Jeżeli TAK, to czy ogólnie rzecz biorąc ten wpływ był raczej korzystny czy raczej niekorzystny?* Na drugie z wymienionych pytań mogli odpowiedzieć posługując się odpowiedziami 1, 2, 4, 5 z tabeli 3.50. Do kategorii 3.*Brak wpływu* zaklasyfikowane zostały osoby, które na pierwsze z wymienionych pytań odpowiedziały „Nie". W analizach zakładamy, że kategorie te tworzą skalę porządkową. Potwierdzenie tego założenia stanowią wyniki logarytmiczno–multiplikatywnego modelu wierszowo–kolumnowego. Parametry skalujące zarówno dla zmiennej wierszowej, jak i kolumnowej dają uporządkowanie zgodne z tym, jakie zostało przyjęte.

Aby oszacować rozkład zgodny z tym modelem metodą największej wiarygodności, poza warunkami dotyczącymi quasi–symetrii[31] (3.19-3.21) musi być spełniony warunek:

$$\sum_{i=1}^{r-1}\sum_{j>i}^{r} i(j-i)\cdot\widehat{\pi}_{ij}^{XY} - \sum_{j=1}^{r-1}\sum_{i>j}^{r} j(i-j)\cdot\widehat{\pi}_{ij}^{XY} = \tag{3.55}$$
$$= \sum_{i=1}^{r-1}\sum_{j>i}^{r} i(j-i)\cdot p_{ij}^{XY} - \sum_{j=1}^{r-1}\sum_{i>j}^{r} j(i-j)\cdot p_{ij}^{XY}.$$

Powyższe wyrażenie zdaje sprawę, z jakimi wartościami jednej zmiennej współwystępują wartości drugiej zmiennej, powyżej i poniżej głównej przekątnej. Rozkład oczekiwany zgodny z modelem $QS\_U\_SK$ pokazuje tablica 3.51.

Tabela 3.51: Rozkład oczekiwany dla danych z tabeli 3.50 zgodny z modelem $QS\_U\_SK$

| Odpowiedzi z 2003 roku | Odpowiedzi z 2005 roku | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Suma |
| 1. Bardzo niekorzystny | 199,6 | 293,8 | 155,4 | 43,5 | 3,8 | 696,3 |
| 2. Raczej niekorzystny | 187,7 | 759,1 | 499,2 | 195,1 | 20,0 | 1661,1 |
| 3. Brak wpływu | 125,3 | 487,3 | 1080,9 | 289,5 | 23,1 | 2006,1 |
| 4. Raczej korzystny | 28,5 | 119,6 | 140,6 | 259,7 | 45,7 | 594,0 |
| 5. Bardzo korzystny | 3,3 | 12,2 | 8,7 | 27,3 | 27,2 | 78,7 |
| Suma | 544,4 | 1672 | 1884,8 | 815,2 | 119,8 | 5036,2 |

Parametr asymetrii wynosi $a = 1,136$. Oznacza to, że wszystkie lokalne stosunki szans znajdujące się powyżej głównej przekątnej są 1,29 ($a^2 = 1,136^2$) razy większe niż stosunki szans położone symetrycznie poniżej przekątnej. Przykładowo:

$$\frac{\Theta_{13}^{XY}}{\Theta_{31}^{XY}} = \frac{(155,4\cdot 195,1)/(43,5\cdot 499,2)}{(125,3\cdot 119,6)/(28,5\cdot 487,3)} = \frac{1,4}{1,08} = 1,29.$$

Oznacza to, że wśród osób, które w 2003 roku uważały, że przemiany po 1989 roku miały na ich życie wpływ *bardzo niekorzystny*, proporcja liczby osób, które w 2005 roku nie zauważyły *żadnego wpływu* do liczby osób, które w 2005 roku oceniły, że był to wpływ *raczej korzystny* była 1,4 razy większa niż analogiczna proporcja dla osób które w 2003 roku oceniły ten wpływ jako *raczej niekorzystny*. Okazuje się, że jeśli wyznaczymy symetryczny stosunek szans to jest on 1,3 razy mniejszy. Tak

[31] Bądź warunkami związanego z innym modelem, który jest modyfikowany przez opisywaną asymetrię.

więc proporcja liczby osób, które w 2005 roku zauważyły wpływ *bardzo niekorzystny* do liczby osób, które w tym samym roku zauważyły wpływ *raczej niekorzystny*, była większa 1,08 (tj.1,4/1,3) razy wśród osób, które w 2003 nie dostrzegły *żadnego wpływu* niż wśród osób, które w tym samych roku dostrzegły wpływ *raczej korzystny*. W podobnych proporcjach pozostają do siebie wszystkie inne lokalne stosunki szans wyznaczone dla komórek ulokowanych symetrycznie względem głównej przekątnej.

Jeśli chodzi o tablicę ruchliwości edukacyjnej, to przypomnijmy, że dla modelu quasi–niezależności statystyka $G^2$ wynosiła 8,11 ($p = 0,0439$) przy 3 stopniach swobody. Uwzględnienie jednakowej asymetrycznej interakcji nie redukuje statystyki $G^2$ w sposób istotny statystycznie: $G^2 = 8,09$ ($p = 0,0175$), $df = 2$, tak więc redukcja statystyki $G^2$ wynosi w przybliżeniu jedynie 0,02.

### Asymetryczne modele poziomu

W modelach tych asymetryczna modyfikacja quasi–symetrii odnosi się do wprowadzonej wcześniej wielkości $\Phi_{ijk}$, porównującej prawdopodobieństwo cyklu $i-j-k-i$ obejmującego dwa przejścia „w górę" oraz jedno przejście „w dół" oraz cyklu $k-j-i-k$, na który składa się jedno przejście „w dół" i dwa przejścia „w górę". Wielkość ta jest równa 1 w modelach quasi–symetrii, natomiast jeśli uwzględnia się asymetrię może być modelowana — jak pokazuje Yamaguchi — za pomocą jednego bądź kilku parametrów. W najprostszej postaci, modyfikacja przybiera postać:

$$\Phi_{ijk} = \frac{F_{ij}^{XY} \cdot F_{jk}^{XY} \cdot F_{ki}^{XY}}{F_{kj}^{XY} \cdot F_{ji}^{XY} \cdot F_{ik}^{XY}} = a^2 \tag{3.56}$$

gdzie $i < j < k$. Modyfikację taką nazywa się *trójkątną asymetrią* — i oznacza się jako jako $TP\_SK$. Ilustrację tego modelu stanowi tabela 3.52. Parametryzacja takiego modelu jest taka sama jak w równaniu 3.52, przy czym parametr asymetrii w modelach typu $TP\_SK$ możemy zdefiniować jako:

$$a_{ij}^{XY} = \begin{cases} a, & \text{dla } i < j \\ \\ \frac{1}{a}, & \text{dla } i > j \end{cases} \tag{3.57}$$

Jak widać *trójkątna* asymetria wymaga jedynie jednego dodatkowego parametru, przy czym komórki powyżej przekątnej modyfikowane są przez wielkość $a$, natomiast komórki poniżej przekątnej przez wartość $1/a$. W podobny sposób modyfikować można inne modele, będące szczególnymi przypadkami quasi–symetrii. Przykładowo, połączenie asymetrii tego typu z modelem quasi–niezależności prowadzi do sformułowania hipotezy, znanej w literaturze (Goodman, 1972) jako model „trójkątny" (*triangel parameter model*). Ilustruje go tablica 3.53.

Tabela 3.52: Ilustracja modelu quasi–symetrii z trójkątną asymetrią — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | a | a | a | a |
| $x_2$ | $\frac{1}{a}$ | $s_{22}$ | $s_{23}a$ | $s_{24}a$ | $s_{25}a$ |
| $x_3$ | $\frac{1}{a}$ | $\frac{s_{23}}{a}$ | $s_{33}$ | $s_{34}a$ | $s_{35}a$ |
| $x_4$ | $\frac{1}{a}$ | $\frac{s_{24}}{a}$ | $\frac{s_{34}}{a}$ | $s_{44}$ | $s_{45}a$ |
| $x_5$ | $\frac{1}{a}$ | $\frac{s_{25}}{a}$ | $\frac{s_{35}}{a}$ | $\frac{s_{45}}{a}$ | $s_{55}$ |

Tabela 3.53: Ilustracja modelu trójkątnego $QN\_TP\_SK$ — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|---|---|---|---|---|---|
| $x_1$ | $q_1$ | $a$ | $a$ | $a$ | $a$ |
| $x_2$ | $\frac{1}{a}$ | $q_2$ | $a$ | $a$ | $a$ |
| $x_3$ | $\frac{1}{a}$ | $\frac{1}{a}$ | $q_3$ | $a$ | $a$ |
| $x_4$ | $\frac{1}{a}$ | $\frac{1}{a}$ | $\frac{1}{a}$ | $q_4$ | $a$ |
| $x_5$ | $\frac{1}{a}$ | $\frac{1}{a}$ | $\frac{1}{a}$ | $\frac{1}{a}$ | $q_5$ |

Aby wyznaczyć rozkład oczekiwany metodą największej wiarygodności zgodny z modelem, w którym uwzględnia się asymetrię „trójkątną", spełniony musi być warunek[32]:

$$\sum_{i=1}^{r-1}\sum_{j>i}^{r}\widehat{\pi}_{ij}^{XY}-\sum_{j=1}^{r-1}\sum_{i>j}^{r}\widehat{\pi}_{ij}^{XY}=\sum_{i=1}^{r-1}\sum_{j>i}^{r}p_{ij}^{XY}-\sum_{j=1}^{r-1}\sum_{i>j}^{r}p_{ij}^{XY} \tag{3.58}$$

Zgodnie z tym warunkiem różnicę pomiędzy prawdopodobieństwami powyżej i poniżej głównej przekątnej szacuje się na podstawie próby.

Warto przypomnieć, że przy opisie tabel o takich samych kategoriach zmiennej wierszowej i kolumnowej na początku tego rozdziału, porównywane były odsetki poniżej i powyżej głównej przekątnej. W odniesieniu do danych panelowych opisujących ocenę wpływu przemian, które zaszły po 1989 roku na życie badanych okazuje się, że 31,1% respondentów oceniało wpływ zmian na własne życie korzystniej w 2005 aniżeli w 2003 roku. Natomiast ocena 22,7% pogorszyła się jeśli porównujemy obydwa punkty czasowe. Różnica odsetków poniżej i powyżej głównej przekątnej $(22,7\% - 31,1\% = -8,4)$ zdaje sprawę z asymetrii, przy czym na ich rozmiar wpływ

[32]Oczywiście spełnione muszą być warunki związane z modelem który jest modyfikowany przez tego typu asymetrię, jeśli jest to model quasi–symetrii są to warunki 3.19-3.21

mają różnice w rozkładach brzegowych. Cytowany wcześniej indeks Liebersona wynosi $-6,4\%$, czyli przy założeniu niezależności taka byłaby różnica pomiędzy odsetkiem osób, których ocena pogorszyła się w stosunku do osób, których ocena się polepszyła.

Aby odpowiedzieć na pytanie, czy różnice te wynikają jedynie z różnic w rozkładach brzegowych warto zastosować modele trójkątnej asymetrii. Okazuje się, że modelu $QS\_TP\_SK$ — tj. modelu quasi-symetrii uwzględniającego trójkątną asymetryczną modyfikację — nie można odrzucić przy standardowo przyjmowanych poziomach istotności: $\chi^2 = 8,07$ $(p = 0,1520)$ oraz $G^2 = 8,10$ $(p = 0,1506)$, $df = 5$. Indeks rozbieżności wynosi $\Delta = 1,1$. Test warunkowy porównujący ten model z modelem quasi–symetrii, pokazuje, że asymetryczna modyfikacja jest istotna na poziomie istotności $0,05$, tj. $G^2 = 17,9 - 8,10 = 9,9$, $(p = 0,0017)$.

W odniesieniu do danych panelowych parametr asymetrii $a$ wynosi 1,17. Jest on większy od 1, co wskazuje, że również wtedy gdy kontrolowane są różnice w rozkładach brzegowych, relatywnie większe są liczebności w komórkach powyżej przekątnej w porównaniu do komórek usytuowanych poniżej. Przy uwzględnieniu (kontroli) tego, że ocena przemian po 1989 roku na życie badanych zmieniła się generalnie (tj. zmienił się rozkład brzegowy), to sytuacja, w której ocena respondenta była bardziej korzystna w 2005 niż 2003 roku, jest $1,37$ (tj. $1,17^2$) razy bardziej prawdopodobna, aniżeli sytuacja odwrotna.

W podobny sposób można interpretować wielkość parametru asymetrii w odniesieniu do analizowanej wcześniej tablicy ruchliwości edukacyjnej. Jest on większy od 1, co oznacza, że kontrolując różnice strukturalne pomiędzy pozycją ojca i syna silniejsza jest tendencja do występowania „awansu" syna w stosunku do pozycji ojca niż tendencja do „degradacji" społecznej. Choć model quasi–symetrii uwzględniający parametr trójkątnej asymetrii jest akceptowalny na poziomie istotności 0,01 ($\chi^2 = 7,26$, $p = 0,0256$ oraz $G^2 = 8,07$, $p = 0,0176$, $df = 2$, $\Delta = 1,0$), to parametr $a = 1,03$ co wskazuje, że asymetria jest niewielka. Co więcej, test warunkowy porównujący modele $QS$ oraz $QS\_TP\_SK$ pokazuje, że uwzględnienie parametru asymetrii nie poprawia dopasowania istotnie statystycznie: $G^2 = 8,11 - 8,07 = 0,04$, $(p = 0,84)$. Można więc przyjąć, że trójkątna asymetria nie jest dobrym opisem procesu ruchliwości w odniesieniu do analizowanych danych.

### Inne modele asymetrii

Powyżej przedstawione zostały najprostsze sposoby uwzględniania asymetrii. Zarówno jednakowa asymetryczna interakcja jak i trójkątna asymetria wymagały uwzględnienia tylko jednego parametru. Yamaguchi przedstawia możliwości formułowania modeli bardziej złożonych, w których asymetria opisywana jest za pomocą kil-

ku parametrów. Przykładowo prezentowany wcześniej asymetryczny model dystansu (D) może być postrzegany jako modyfikacja symetrycznego modelu dystansu (DS) przez uwzględnienie parametrów asymetrii specyficznych dla kolejnych pseudo–przekątnych. W pracy tej nie będą przedstawiane inne modele, które wymagają uwzględnienia kilku parametrów. Na ogół ich interpretacja nie jest łatwa. Testowanie różnych złożonych typów asymetrii bez należytego namysłu teoretycznego - nawet jeśli prowadzi do wyboru modelu dobrze opisującego dane - nie musi być strategią sensowną. Dobre dopasowanie może być do pewnego stopnia przypadkowe, w tym sensie, że nawet jeśli liczebności oczekiwane nie będą odbiegać znacząco od obserwowanych, nie musi to oznaczać, że model trafnie odzwierciedla opisywane zjawisko.

Warto ponadto zwrócić uwagę, że asymetryczna modyfikacja może być uwzględniana w odniesieniu do różnych quasi–symetrycznych modeli. Powyżej zasygnalizowane zostały jedynie wybrane kombinacje modeli quasi–symetrycznych z różnymi asymetrycznymi, tj. $QS\_U\_SK$, $UA\_U\_SK$, $QS\_TP\_SK$, $QN\_TP\_SK$. Możliwości tych jest znacznie więcej. Trzeba jednak pamiętać, że formułując hipotezy asymetryczne istotne jest na ile model w całości wydaje się sensowny. Przykładowo, na ile interpretacja parametrów asymetrycznych jest spójna z interpretacją parametrów symetrycznych.

### Specyficzne asymetryczne modele dla dwóch zmiennych

Powyższe analizy tablicy ruchliwości edukacyjnej pokazały, że ani uwzględnienie jednakowej asymetrycznej interakcji ani uwzględnienie trójkątnej asymetrii nie poprawia istotnie dopasowania modelu quasi–symetrii. Można zadać pytanie, jakiego typu modyfikacja modelu byłaby w tej sytuacji adekwatna?

Często stosowaną metodą jest porównanie liczebności oczekiwanych z liczebnościami empirycznymi. Warto jednak zauważyć, że różnice te są silnie związane z liczebnością danej komórki. Dlatego też, do oceny rozbieżności liczebności danej komórki stosuje się tzw. *reszty Pearsona*[33] (*Pearsonian residuals*), tj. wyrażenia:

$$e_{ij}^{XY} = \frac{f_{ij}^{XY} - \widehat{F}_{ij}^{XY}}{\sqrt{\widehat{F}_{ij}^{XY}}}, \tag{3.59}$$

gdzie $f_{ij}^{XY}$ są liczebnościami z próby, $\widehat{F}_{ij}^{XY}$ liczebnościami zgodnymi z modelem. Wyrażenie w mianowniku tj. $\sqrt{\widehat{F}_{ij}^{XY}}$ może być przy pewnych założeniach interpretowane

[33]Lepsze własności mają tzw. standaryzowane reszty (Haberman 1973, 1978, Agresti 1984, 2002, 2007), dokładne omówienie różnic pomiędzy tymi wielkościami przekracza jednak ramy tej pracy. W tym miejscu chodzi jedynie o zasygnalizowanie, że istnieją miary, które porównują dopasowanie do modelu konkretnych komórek.

jako odchylenie standardowe różnicy $f_{ij}^{XY} - \widehat{F}_{ij}^{XY}$, czyli zdaje sprawę na ile — przy założeniu prawdziwości rozpatrywanego modelu — w poszczególnych próbach liczebności danej kombinacji obydwu zmiennych różnią się od liczebności szacowanej w modelu. Zauważmy, że statystyka $\chi^2$ zdefiniowana w formule 1.63 opiera się na sumie kwadratów reszt (tj. $e_{ij}^2$) dla wszystkich komórek.

W tabeli 3.54 przedstawiono reszty Pearsona obliczone dla modelu quasi–symetrii w odniesieniu do tablicy ruchliwości edukacyjnej. Analiza tych wielkości wskazuje, że stosunkowo największa rozbieżność dotyczy liczebności $f_{42}^{XY}$, tj. respondentów posiadających wykształcenie niepełne średnie, których ojcowie posiadali wykształcenie wyższe. Wielkość dodatnia reszty wskazuje, że osoby takie występują w próbie częściej niż wskazuje na to model QS.

Tabela 3.54: Reszty dla modelu QS dla tablicy ruchliwości edukacyjnej 3.5

| | Wykształcenie syna | | | |
|---|---|---|---|---|
| Wykształcenie ojca | 1 | 2 | 3 | 4 |
| 1. Podstawowe i niepełne podstawowe | 0,000 | -0,157 | 0,190 | 0,087 |
| 2. Niepełne średnie (w tym zasadnicze zawodowe) | 0,719 | 0,000 | 0,162 | -0,761 |
| 3. Ukończone średnie | -1,031 | -0,193 | 0,000 | 0,496 |
| 4. Niepełne wyższe i wyższe | -0,951 | 1,751 | -0,963 | 0,000 |

Możliwa jest modyfikacja modelu quasi–symetrii w taki sposób, żeby uwzględnić specyfikę tej kategorii podobnie jak czyni się z komórkami leżącymi na głównej przekątnej np. w modelu quasi–niezależności. Można więc podzielić zbiorowość na dwie podzbiorowości $Z = 1$ oraz $Z = 2$. Dla pierwszej z nich zachodzić będzie model quasi–symetrii, druga podzbiorowość, obejmuje osoby, dla których $X = 4$ i $Y = 2$.

Model taki w stosunku do quasi–symetrii posiada tylko jeden parametr więcej, związany z prawdopodobieństwem $\pi_{42}^{XY}$. Okazuje się, że bardzo dobrze odzwierciedla on dane : $\chi^2 = 1,86$, $(p = 0,3930)$ oraz $G^2 = 2,53$, $(p = 0,2816)$, $df = 2$, $\Delta = 0,05$. Test warunkowy porównujący ten model z quasi–symetrią pokazuje, że uwzględnienie dodatkowego parametru jest istotne statystycznie na standardowo przyjmowanym poziomie istotności 0,05: $G^2 = 5,57$, $p = (0,0183)$, $df = 1$. Parametr który wprowadziliśmy do tego modelu wynosi $q_{42}^{XY} = 2,98$, co wskazuje, że respondenci o wy-

kształceniu niepełnym średnim, których ojcowie mają wykształcenie wyższe zdarzają się prawie trzykrotnie częściej niż wynika z modelu quasi–symetrii.

Należy jednak podkreślić, że powyższa strategia modyfikacji modelu w celu poprawy jego dopasowania jest ryzykowna. Jeśli wprowadza się jakiś parametr do modelu powinno się mieć przesłanki teoretyczne, tj. dysponować uzasadnieniem, dlaczego oczekuje się, że dana kombinacja będzie występowała relatywnie częściej (bądź rzadziej). Słabe dopasowanie modelu może mieć przyczyny losowe. Celem jest sformułowanie modelu, który będzie trafnie opisywał dane zjawisko a nie konkretne dane z próby. Jeśli opisana powyżej procedura analizy reszt skłania nas do modyfikacji modelu, warto skonfrontować tak zmodyfikowany model z innymi danymi aby przekonać się czy zaobserwowane zależności nie wynikają jedynie ze specyfiki konkretnego badania.

## 3.4 Modele dla większej liczby zmiennych

Modele zaprezentowane powyżej dotyczyły sytuacji, gdy analizowany był związek dwóch zmiennych o takich samych kategoriach. W tej części zasygnalizowane zostaną możliwości formułowania modeli, w których uwzględnia się kolejne zmienne. Trzeba jednak na wstępie zaznaczyć, że w zależności od charakteru kolejnych zmiennych można stawiać inne pytania badawcze i w konsekwencji mamy do czynienia z modelami różnego typu.

Z jednej strony trzecia zmienna może definiować podzbiorowości, ze względu na które analizowany jest związek dotyczący dwóch zmiennych o takich samych kategoriach. Przykładowo można porównywać, jaki jest związek pomiędzy tą samą cechą mierzoną w kolejnych turach badania panelowego wśród kobiet i mężczyzn. Podobnie można porównywać ruchliwość w różnych punktach czasowych bądź różnych krajach. W wymienionych przykładach trzecia zmienna (płeć, rok badania, kraj) ma charakter zmiennej grupującej. Oczywiście możliwe jest definiowanie podzbiorowości za pomocą kilku zmiennych jednocześnie. Przykładowo, można porównywać ruchliwość społeczną uwzględniając jednocześnie wiek badanych i kraj z którego pochodzą, czyli np. porównywać na ile podobne są wzory ruchliwości wśród młodych Polaków i starszych Holendrów.

Inne modele będą dotyczyć sytuacji, gdy kilka zmiennych — oznaczmy je $X_1$, $X_2$, $X_3$ — ma identyczne kategorie. Na przykład, można pytać respondentów o opinie dotyczącą tej samej kwestii w roku 1995, 2000, i 2005. Można wówczas pytać o siłę związku pomiędzy parami zmiennych, przykładowo, czy związek pomiędzy zmiennymi

$X_1$ oraz $X_2$ jest silniejszy aniżeli pomiędzy zmiennymi $X_1$ i $X_3$, ewentualnie czy związek pomiędzy $X_2$ oraz $X_3$ zależy od wartości zmiennej $X_1$.

Ostatni z wyróżnionych w tym rozdziale przypadków dotyczy sytuacji, gdy mierzymy te same dwie zmienne niekoniecznie o takich samych kategoriach— dajmy na to $X$ oraz $Y$ — w dwóch punktach czasowych. Mamy wówczas do czynienia z czterema zmiennymi, które oznaczyć możemy przykładowo $X_1$, $X_2$, $Y_1$, $Y_2$. Zauważmy, że wówczas zmienne $X_1$ oraz $X_2$ mają te same kategorie, podobnie takie same kategorie mają zmienne $Y_1$ i $Y_2$. Interesująca może być odpowiedź na pytanie czy związek pomiędzy obydwiema zmiennymi był taki sam w obydwu badaniach, a jeśli się zmienił to w jakim stopniu. Innymi słowy odpowiadamy na pytanie czy związek pomiędzy zmiennymi $X_1$ oraz $Y_1$ jest taki sam, czy też inny jak pomiędzy zmiennymi $X_2$ i $Y_2$. Można zadawać również inne pytania, przykładowo: czy tendencja do współwystępowania tej samej kategorii zmiennej $Y$ (tj. sytuacji, gdy $Y_1 = Y_2$) zależy od wartości $X_1$. Modele tego typu nie muszą dotyczyć wyłącznie danych panelowych. Tablice o podobnej strukturze uzyskamy, jeśli zestawimy ze sobą dwie cechy respondenta z tymi samymi cechami określonymi dla jego ojca.

Szczegółowe omówienie modeli związanych z opisanymi powyżej sytuacjami wykracza poza ramy tej pracy[34]. Poniżej zasygnalizowane zostaną pewne możliwości formułowania modeli, które wydają się adekwatne do analizy danych tego typu. Ich prezentacja będzie oparta na przykładowych analizach danych empirycznych.

### 3.4.1 Modelowanie warunkowej zależności dla tablic ruchliwości

Pierwszy z przykładów dotyczy analizy związku pomiędzy dwiema zmiennymi o identycznych kategoriach w podzbiorowościach wyróżnionych ze względu na kategorie trzeciej zmiennej. Ilustrację danych tego typu stanowi tabela 3.55. Pochodzą one z pierwszej edycji Europejskiego Sondażu Społecznego (2002). Tabela dotyczy tablic ruchliwości edukacyjnej dla czterech wybranych krajów: Holandii, Polski, Słowenii i Szwecji. Kraje te są zróżnicowane pod wieloma względami (m. in. historycznie, kulturowo, i gospodarczo) i w tym sensie interesująca wydaje się odpowiedź na pytanie czy różnice te przekładają się na bariery społeczne dotyczące uzyskiwania wykształcenia związane z pochodzeniem. Analiza objęła mężczyzn w wieku 25-54 lata. Zarówno pytanie o wykształcenie respondenta jak i jego ojca zostało zakodowane zgodnie z międzynarodowym standardem ISCED. Przypomnijmy, że klasyfikacja ta definiuje 7 kategorii, aby liczebności były dostatecznie duże konieczne było połączenie wybra-

[34]Ich bardziej szczegółowe omówienie czytelnik znajdzie w (Fingleton 1984).

nych kategorii[35] i w ten sposób wyodrębnione zostały cztery kategorie wymienione w tabeli.

W analizie danych tego typu jako punkt wyjścia przyjmuje się na ogół model warunkowej niezależności, przedstawiony w rozdziale pierwszym, tj. $[XZ][YZ]$ gdzie $X$ to wykształcenie ojca, $Y$ to wykształcenie respondenta a zmienna $Z$ wskazuje na kraj. Na ogół zakłada się, że struktura wykształcenia — zarówno w odniesieniu do pokolenia ojców jak też respondentów — może się różnić w poszczególnych krajach, stąd też nie rozpatruje się modeli prostszych. Hipoteza głosi natomiast, że w każdym kraju wykształcenie respondenta nie zależy od wykształcenia ojca.

Jak łatwo się domyślić model ten nie jest realistycznym opisem danych. Potwierdzają to wyniki weryfikacji przedstawione w tabeli 3.56. Wskazuje to na istnienie zależności pomiędzy wykształceniem ojca i syna. Można zapytać, czy ta zależność jest taka sama, czy też inna w poszczególnych krajach. W tym celu można przywołać hipotezę — sformułowaną również w rozdziale 1 — o identyczności warunkowych stosunków szans, tj. $[XY][YZ][XZ]$. Przypomnijmy, że hipoteza ta głosiła, że zarówno *wzór*, jak też *siła* zależności — opisywane za pomocą stosunków szans pomiędzy zmienną $X$ i $Y$ - są takie same dla każdej podzbiorowości wyróżnionej przez zmienną $Z$. W analizowanym przypadku oznacza to, że jeśli porównuje się poszczególne kraje to wzór i siła ruchliwości edukacyjnej pozostają takie same. Jak pokazuje tabela 3.56 model ten również powinien być odrzucony na poziomie istotności 0,01. Wynik ten wskazuje, że związek pomiędzy wykształceniem ojca i syna różni się w krajach objętych analizą.

Kolejna hipoteza, którą można sformułować w odniesieniu do tabeli 3.55 to model warstwowy LM zaproponowany przez Xie (1992). Hipoteza ta została szczegółowo przedstawiona w rozdziale drugim. Przypomnijmy, że zgodnie z tym modelem o ile wzór zależności pomiędzy zmiennymi $X$ i $Y$ jest w poszczególnych krajach taki sam, to jej siła jest różna. Dokładniej opisuje to formuła 2.111.

Model ten — jak pokazuje tabela 3.56 — jest akceptowalny na poziomie istotności 0,01. Prześledźmy wybrane parametry opisujące wzór zależności oraz jej siłę z formuły 2.108, tj.

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \Psi_{ij}^{\Upsilon_k}.$$

Parametry zostały wyskalowane w ten sposób, że wartości $\Psi_{ij}$, opisujące interakcję pomiędzy wykształceniem ojca i wykształceniem syna przedstawione w tabeli 3.57 dotyczą Holandii[36]. Dlatego wartość parametru $\Upsilon_k$ opisującego siłę zależności dla

[35]Porównaj informacje podane przy okazji omówienia tabeli 2.27.

[36]Przyjęto przy ich wyznaczaniu parametryzację odchyleń multiplikatywnych przedstawioną w pierwszym rozdziale, więc ich iloczyn w każdym wierszu i w każdej kolumnie jest równy 1.

Tabela 3.55: Tablice ruchliwości edukacyjnej dla czterech krajów europejskich[a]

| Holandia | | | | |
|---|---|---|---|---|
| Wykształcenie ojca \ syna | 1 | 2 | 3 | 4 |
| 1. Podstawowe i niepełne podstawowe | 16,7 | 64,2 | 52,9 | 28,9 |
| 2. Niepełne średnie (w tym zasadnicze zawodowe) | 0,5 | 66,6 | 75,9 | 75,9 |
| 3. Ukończone średnie | 3,9 | 11,8 | 33,8 | 54,9 |
| 4. Niepełne wyższe i wyższe | 1,5 | 8,8 | 26,5 | 57,3 |
| Polska | | | | |
| Wykształcenie ojca \ syna | 1 | 2 | 3 | 4 |
| 1. Podstawowe i niepełne podstawowe | 60,3 | 137,3 | 46,6 | 28,8 |
| 2. Niepełne średnie (w tym zasadnicze zawodowe) | 12,1 | 77,1 | 43,4 | 27,1 |
| 3. Ukończone średnie | 1,8 | 15,1 | 36,4 | 30,1 |
| 4. Niepełne wyższe i wyższe | 0,0 | 1,7 | 2,9 | 25,7 |
| Słowenia | | | | |
| Wykształcenie ojca \ syna | 1 | 2 | 3 | 4 |
| 1. Podstawowe i niepełne podstawowe | 52,0 | 58,0 | 30,0 | 13,0 |
| 2. Niepełne średnie (w tym zasadnicze zawodowe) | 9,0 | 50,0 | 48,0 | 28,0 |
| 3. Ukończone średnie | 5,0 | 13,0 | 24,0 | 24,0 |
| 4. Niepełne wyższe i wyższe | 2,0 | 5,0 | 13,0 | 18,0 |
| Szwecja | | | | |
| Wykształcenie ojca \ syna | 1 | 2 | 3 | 4 |
| 1. Podstawowe i niepełne podstawowe | 57,0 | 94,0 | 93,0 | 68,0 |
| 2. Niepełne średnie (w tym zasadnicze zawodowe) | 6,0 | 20,0 | 22,0 | 32,0 |
| 3. Ukończone średnie | 3,0 | 6,0 | 13,0 | 13,0 |
| 4. Niepełne wyższe i wyższe | 2,0 | 13,0 | 18,0 | 75,0 |

[a]Źródło: Europejski Sondaż Społeczny, 2002-2003, dane przeważone.

Tabela 3.56: Wyniki weryfikacji hipotez dla tabeli 3.55

| Model | df | $\chi^2$ | $L^2$ | $\Delta$ |
|---|---|---|---|---|
| $[XZ][YZ]$ | 36 | 440,9 ($p < 0,0001$) | 436,9 ($p < 0,0001$) | 17,4 |
| $[XZ][YZ][XY]$ | 27 | 45,5 ($p = 0,0144$) | 47,7 ($p = 0,0084$) | 4,4 |
| $LM$ | 24 | 35,1 ($p = 0,0674$) | 37,3 ($p = 0,0406$) | 4,2 |
| $DS$ | 33 | 65,6 ($p = 0,0006$) | 64,9 ($p = 0,0008$) | 5,5 |
| $QDS$ | 30 | 66,4 ($p = 0,0001$) | 62,1 ($p = 0,0005$) | 5,0 |
| $QDSA$ | 29 | 48,8 ($p = 0,0122$) | 49,8 ($p = 0,0095$) | 4,5 |
| $QDS_Z$ | 21 | 36,4 ($p = 0,0195$) | 36,9 ($p = 0,0171$) | 3,4 |
| $Q_ZDS_Z$ | 12 | 28,2 ($p = 0,0051$) | 22,8 ($p = 0,0292$) | 1,5 |
| $DS_ZA$ | 23 | 26,5 ($p = 0,2769$) | 30,2 ($p = 0,1451$) | 3,1 |
| $Q_ZDS_ZA$ | 11 | 15,9 ($p = 0,1436$) | 15,6 ($p = 0,1548$) | 1,4 |
| $Q_ZDS_ZA_Z$ | 8 | 12,6 ($p = 0,1270$) | 13,9 ($p = 0,0855$) | 1,2 |

tego kraju wynosi 1, a wartości dla pozostałych krajów zostały ustalone względem niej i wynoszą odpowiednio: dla Polski parametr wynosi 1,37, dla Słowenii 1,11, dla Szwecji 0,82.

Tabela 3.57: Parametry interakcji $\Psi_{ij}$ z modelu warstwowego dla tabeli 3.55

| Wykształcenie ojca \ syna | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. Podstawowe i niepełne podstawowe | 2,71 | 1,39 | 0,71 | 0,37 |
| 2. Niepełne średnie (w tym zasadnicze zawodowe) | 0,86 | 1,43 | 1,03 | 0,79 |
| 3. Ukończone średnie | 0,83 | 0,73 | 1,28 | 1,29 |
| 4. Niepełne wyższe i wyższe | 0,51 | 0,69 | 1,08 | 2,63 |

Na podstawie podanych powyżej parametrów możliwe jest zrekonstruowanie dowolnego stosunku szans opisującego siłę zależności pomiędzy kategoriami wykształcenia respondenta i jego ojca dla dowolnego kraju. Przykładowo wartość lokalnego stosunku szans $\Theta_{11}^{XY}$ dla Holandii wynosi:

$$\Theta_{11(1)}^{XY(Z)} = \frac{2,71 \cdot 1,43}{1,39 \cdot 0,86} = 3,21.$$

Dla Polski wartość ta wynosi:

$$\Theta_{11(2)}^{XY(Z)} = \left(\Theta_{11(1)}^{XY(Z)}\right)^{\Upsilon_2} = 3,21^{1,37} = 4,93.$$

W takiej samej relacji pozostają wszystkie pozostałe stosunki szans np.

$$\Theta_{12(2)}^{XY(Z)} = \left(\Theta_{12(1)}^{XY(Z)}\right)^{\Upsilon_2}$$

itd. Jeśli przyjąć, że większa siła zależności pomiędzy wykształceniem respondenta i jego ojca wskazuje na większe bariery edukacyjne związane z dziedziczeniem pozycji, to najbardziej „otwartym" pod tym względem krajem jest Szwecja, gdyż dla tego kraju wartość parametru $\Upsilon_k$ jest najmniejsza. Następnie należałoby wymienić Holandię, natomiast w Słowenii i szczególnie w Polsce wykształcenie respondenta jest stosunkowo silnie związane z wykształceniem ojca.

Przytoczone powyżej hipotezy dotyczyły jedynie tego, czy istnieje zależność pomiędzy zmiennymi $X$ i $Y$, jak również czy siła i wzór tej zależności są takie same czy też inne w poszczególnych krajach. Możliwe jest jednak modelowanie wzoru zależności w sposób prostszy. Można na przykład założyć, że związek pomiędzy wykształceniem ojca i wykształceniem syna daje się opisać za pomocą modelu symetrii. Jeśli uwzględni się informację o porządkowym charakterze tych zmiennych, to dodatkowo można wykorzystać model jednakowej interakcji, modele wierszowo–kolumnowe, itd. Dodatkowe możliwości wynikają z tego, że kategorie zmiennych $X$ i $Y$ są identyczne. Nie będą formułowane wszystkie modele tego typu, prześledzone zostaną jedynie możliwości wykorzystania do tego celu symetrycznej wersji modelu dystansu. Przykładowo, jeśli założymy, że w każdym kraju ruchliwość edukacyjną można opisać za pomocą modelu symetrycznego dystansu, wówczas warunkowe stosunki szans zmiennych $X$ i $Y$ względem $Z$ można przedstawić analogicznie do warunków 3.38 oraz 3.39:

$$\Theta_{i(i+m)(k)}^{XY\quad(Z)} = \Theta_{(i+m)i(k)}^{X\quad Y(Z)} = \Theta_m \tag{3.60}$$

gdzie $m = |i - j|$. Parametryzacja tego modelu przedstawia się następująco:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \delta_m \tag{3.61}$$

Zauważmy, że w modelu tym zakłada się, że zarówno wzór jak i siła zależności jest taka sama w każdym kraju: zarówno wielkość $\Theta_m$ jak i parametr $\delta_m$ są identyczne dla każdego kraju, tj. nie są one indeksowane przez wartości zmiennej Z. Zależność ta jest wiec opisana prościej niż w modelu niż w modelach $[XY][YZ][XZ]$ oraz LM. Do jej opisu wystarcza $r - 1$ parametrów specyficznych dla kolejnych pseudo–przekątnych. Model ten — będziemy go oznaczać DS — posiada zaledwie o $r - 1$ parametrów więcej niż model warunkowej niezależności.

Jak pokazują wyniki weryfikacji przedstawione w tabeli 3.56, model taki musi być odrzucony na standardowo przyjmowanych poziomach istotności. Słabe dopasowanie modelu może wynikać z tego, że nie zdaje on sprawy ze specyfiki poszczególnych komórek na głównej przekątnej. Choć w modelu tym uwzględnia się — przez kontrast z innymi pseudo–przekątnymi — że komórki na głównej przekątnej występują relatywnie częściej, niemniej zakłada się w nim, że tendencja ta — przy ustalonych rozkładach brzegowych — jest tak samo silna dla każdej kategorii zmiennej. Tym samym nie uwzględnia się, że tendencja dziedziczenia wykształcenia może być silniejsza — przykładowo — dla wykształcenia wyższego niż dla wykształcenia średniego. Można uwzględnić dodatkowo $r-1$ parametrów różnicujących komórki na głównej przekątnej, przy czym wartości te będą takie same dla każdego kraju. Tak sformułowany model oznaczymy QDS. Modyfikacja ta nieznacznie poprawia dopasowanie modelu do danych, niemniej na poziomie 0,01 należałoby go odrzucić.

Możliwe jest również uwzględnienie asymetrii. Przykładowo można uwzględnić trójkątną asymetrię, tj. założyć, że komórki powyżej przekątnej są relatywnie bardziej prawdopodobne, jeśli kontrolujemy rozbieżności w rozkładach brzegowych. Jeśli założymy, że siła asymetrii jest taka sama dla każdego kraju, wówczas modyfikacja taka wiąże się z uwzględnieniem tylko jednego parametru. Model taki — oznaczony w tabeli 3.56 jako QDSA — również musiałby zostać odrzucony na poziomie istotności 0,01.

Tak sformułowany model posiada trzy parametry opisujące kolejne pseudo–przekątne, trzy parametry różnicujące główną przekątną, jak również jeden parametr asymetrii. Parametry te są jednak takie same dla wszystkich krajów, i dlatego też jest on szczególnym przypadkiem hipotezy $[XY][XZ][YZ]$. Założenia te można jednak osłabić. Przykładowo, można uwzględnić, że parametry poszczególnych pseudo–przekątnych są specyficzne dla każdego kraju. Podobnie zróżnicowane mogą być parametry związane z główną przekątną i parametry asymetrii. Model, który uwzględnia zróżnicowanie wszystkich parametrów można przedstawić w sposób następujący:

$$\pi_{ijk}^{XYZ} = \gamma \cdot \gamma_i^X \cdot \gamma_j^Y \cdot \gamma_k^Z \cdot \gamma_{ik}^{XZ} \cdot \gamma_{jk}^{YZ} \cdot \delta_{mk} \cdot q_{ik} \cdot a_k \tag{3.62}$$

gdzie $m = |i-j|$, $q_i = 1$, jeśli $i \neq j$, natomiast $a = b$, gdy $i < j$ oraz $a = 1/b$ gdy $i > j$. Model taki został oznaczony $Q_Z DS_Z A_Z$. Indeksy dolne $Z$ wskazują, że parametry głównej przekątnej (na co wskazuje symbol Q), pseudo–przekątnych (DS) jak też asymetrii (A), mogą się różnić w poszczególnych krajach. Jak widać z tabeli $Q_Z DS_Z A_Z$ model ten jest akceptowalny na poziomie istotności 0,05.

Parametry tego modelu zostały przedstawione w tabeli 3.58. Dla każdego kraju można zaobserwować, że $d_1 > d_2 > d_3$. Wskazuje to, że przy kontroli różnicy w

Tabela 3.58: Parametry modelu $Q_Z DS_Z A_Z$

| Kraj | Parametry dystansu | | | Parametry przekątnej | | | Asymetria |
|---|---|---|---|---|---|---|---|
| | $d_1$ | $d_2$ | $d_3$ | $q_1$ | $q_2$ | $q_3$ | $b$ |
| Holandia | 0,948 | 0,660 | 0,260 | 2,464 | 1,842 | 1,283 | 1,715 |
| Polska | 0,234 | 0,089 | 0,033 | 0,366 | 0,427 | 0,547 | 1,144 |
| Słowenia | 0,880 | 0,518 | 0,243 | 2,531 | 1,321 | 1,151 | 1,538 |
| Szwecja | 0,412 | 0,361 | 0,156 | 0,681 | 0,463 | 0,700 | 1,065 |

rozkładach brzegowych relatywnie rzadziej — w porównaniu do osób niemobilnych — zdarzają się osoby, których wykształcenie znacznie różni się od wykształcenia ich ojca. Ciekawe jest porównanie wielkości tych parametrów między krajami. Dla Holandii parametr $d_1$ ma znacznie wyższą wartość niż ten sam parametr dla Polski, co więcej: również parametr $d_3$ dla Holandii ma nieco wyższą wartość niż parametr $d_1$ dla Polski. Wynik ten potwierdza, że bariery ruchliwości edukacyjnej w Polsce są większe niż w Holandii, co więcej podobna konkluzja dotyczy porównania Polski do innych krajów.

Powyższe parametry są symetryczne, tj. w porównaniu do modelu dystansu $d_1 = d_{(-1)}$. Gdyby parametr asymetrii był równy 1, wielkość $\Phi_{ijl}$, porównująca prawdopodobieństwo wystąpienia dwóch cykli $i-j-l-i$ oraz $l-j-i-l$, musiałaby być równa 1, dla dowolnych wartości $i$, $j$, $l$ zmiennych $X$ i $Y$. Założenie, to nie koniecznie jest trafne. Parametry asymetrii są dla każdego kraju większe od 1 co wskazuje, że kontrolując różnice w rozkładach brzegowych relatywnie częściej mamy do czynienia z poprawą a nie pogorszeniem wykształcenia respondenta względem wykształcenia jego ojca. O ile jednak w Szwecji parametr ten jest bliski jest wartości 1, co jest bliskie symetrii, to w Holandii jego wartość przekracza 1,7.

Parametry $q_1$, $q_2$, $q_3$ wskazują na zróżnicowanie siły dziedziczenia w poszczególnych kategoriach wykształcenia. Parametr dla wykształcenia wyższego ma wartość 1 i względem tej wielkości zostały ustalone pozostałe wartości: $q_1$ dla wykształcenia podstawowego, $q_2$ dla niepełnego średniego oraz $q_3$ dla kategorii „ukończone średnie". Widać pewne różnice między krajami: w Holandii i Słowenii najsilniejsza jest tendencja do dziedziczenia wykształcenia podstawowego, w Polsce i Szwecji wszystkie parametry są mniejsze od 1, co wskazuje że wykształcenie wyższe jest dziedziczone relatywnie najczęściej.

Należy zwrócić uwagę, że możliwe jest sformułowanie wielu modeli, w zależności od tego, które parametry opisujące związek między $X$ i $Y$ zostaną uwzględnione,

bądź co do których z nich dopuścimy zróżnicowanie ze względu na wartości zmiennej grupującej $Z$. Niektóre z nich przedstawione zostały w tabeli 3.56. Przykładowo model $DS_ZA$ głosi, że parametry opisujące poszczególne pseudo–przekątne różnią się w poszczególnych krajach, asymetria jest taka sama, ponadto nie się uwzględnia zróżnicowania jeśli chodzi o siłę dziedziczenia dla poszczególnych kategorii wykształcenia. Model ten jest prostszy od poprzednio omawianego, a jego dopasowanie jest zadowalające. Testy warunkowe potwierdzają, że obydwa te założenia nie pogarszają dopasowana modelu w sposób istotny statystycznie.

Jak widać dla dwóch zmiennych o takich samych kategoriach analizowanych w różnych podzbiorowościach — w tym przypadku tablic ruchliwości porównywanych w różnych krajach — można sformułować wiele modeli w zależności od tego, czy wielkości poszczególnych parametrów są takie same, czy też różne w poszczególnych krajach. Oczywiście, ruchliwość może być opisywana za pomocą innego procesu, aniżeli ten, który wynika z modelu dystansu. Powyższy przykład sygnalizuje jedynie możliwości modelowania warunkowej zależności. Liczba różnych modeli, które daje się sformułować w ten sposób jest znacząca. Oczywiście, trzeba wziąć pod uwagę, że połączenie parametrów — opisujących typ związku, główną przekątną, asymetrię — powinno być uzasadnione merytorycznie.

### 3.4.2 Analiza rozkładu trzech zmiennych o takich samych kategoriach

W tej części przedstawione zostaną metody analizy tablic, w których więcej niż dwie zmienne mają takie same kategorie. Przykładem danych tego typu jest rozkład przedstawiony w tabeli 3.59. Prezentują one dane panelowe pochodzące z kolejnych edycji badania Diagnoza Społeczna, tj. z 2000, 2003 i 2005 roku. Respondentów zapytano trzykrotnie o to, jak często w ostatnich miesiącach, problemy i kłopoty finansowe przysparzały im zmartwień i utrudniały życie.[37]

W tabeli 3.60 przedstawione zostały wybrane modele jakie można sformułować w odniesieniu do danych tego typu. Okazuje się, że modele, w których jedna lub więcej par zmiennych pozostaje warunkowo niezależna nie są akceptowalne na standardowo przyjmowanym poziomie istotności. Przykładowo, w modelu $[XY][YZ]$ zakłada się, że opinie w 2000 i 2005 roku — tj. najbardziej odległych punktach czasowych — pozostają niezależne stochastycznie w podzbiorowościach wyróżnionych, ze względu

[37] W 2000 roku do treści pytania dodano sformułowanie „bardziej niż zazwyczaj", w tym sensie pytania nie były identyczne, niemniej wystarczająco podobne by przykład ten ilustrował modele, jakie można zastosować do analizy rozkładu trzech zmiennych o takich samych kategoriach.

Tabela 3.59: Odpowiedzi na pytanie dotyczące kłopotów związanych z sytuacją finansową w kolejnych latach[a], X–2000, Y–2003, Z–2005 rok.

| W ostatnich miesiącach problemy i kłopoty finansowe przysparzały zmartwień i utrudniały Panu(i) życie: | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Często($Z=1$) | | | Zdarzyło się($Z=2$) | | | Nigdy($Z=3$) | | |
| X\ Y | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 1 | 369,1 | 193,8 | 15,7 | 201,3 | 209,9 | 27,9 | 30,2 | 23,9 | 10,7 |
| 2 | 125,8 | 236,3 | 46,4 | 149,5 | 625,6 | 140,2 | 31,4 | 139,9 | 68,3 |
| 3 | 8,4 | 20,0 | 14,0 | 24,6 | 108,2 | 76,2 | 9,2 | 60,5 | 124,8 |

[a]Źródło: Diagnoza Społeczna, 2000–2005.

Tabela 3.60: Wyniki weryfikacji modeli dla danych z tabeli 3.59

| Model | df | $\chi^2$ | $L^2$ | $\Delta$ |
|---|---|---|---|---|
| $[XY][XZ]$ | 12 | 509,4 ($p < 0,0001$) | 465,1 ($p < 0,0001$) | 1,3 |
| $[XY][YZ]$ | 12 | 163,6 ($p < 0,0001$) | 152,0 ($p < 0,0001$) | 8,9 |
| $[XZ][YZ]$ | 12 | 274,1 ($p < 0,0001$) | 266,6 ($p < 0,0001$) | 10,9 |
| $[XY][XZ][YZ]$ | 8 | 5,7 ($p = 0,6821$) | 5,8 ($p = 0,6696$) | 1,3 |
| $Q[XY][XZ][YZ]$ | 5 | 4,3 ($p = 0,5028$) | 4,4 ($p = 0,4954$) | 0,8 |
| $[XY_{UA}][XZ_{UA}][YZ_{UA}]$ | 17 | 77,0 ($p < 0,0001$) | 70,6 ($p < 0,0001$) | 5,9 |
| $[XY_{QhUA}][XZ_{QhUA}][YZ_{QhUA}]$ | 14 | 14,5 ($p = 0,4121$) | 14,4 ($p = 0,4192$) | 2,1 |
| $Q[XY_{QhUA}][XZ_{QhUA}][YZ_{QhUA}]$ | 11 | 11,2 ($p = 0,4271$) | 11,1 ($p = 0,4326$) | 1,5 |
| $[XY_{UA}][XZ_{QhUA}][YZ_{QhUA}]$ | 15 | 18,4 ($p = 0,2414$) | 18,1 ($p = 0,2600$) | 2,4 |
| $[XY_{QhUA}][XZ_{UA}][YZ_{QhUA}]$ | 15 | 32,4 ($p = 0,0058$) | 31,6 ($p = 0,0072$) | 3,6 |
| $[XY_{QhUA}][XZ_{QhUA}][YZ_{UA}]$ | 15 | 44,9 ($p < 0,0001$) | 41,1 ($p = 0,0003$) | 4,1 |

na to jaką odpowiedź respondenci wskazali w roku 2003. Model taki musimy odrzucić na poziomie 0,01. Podobnie modele $[XY][XZ]$ oraz $[XZ][YZ]$.

W związku z tym nie wydaje się sensowne formułowanie modeli prostszych, w których niezależność dotyczyłaby dwóch par zmiennych bądź wszystkich trzech zmiennych. Bardziej złożony model $[XY][XZ][YZ]$ głosi identyczność warunkowych stosunków szans. Zakłada się w nim, że związek pomiędzy odpowiedziami w roku 2003 i 2005 nie zależy od tego, co deklarował respondent w 2000 roku. Model ten jest ak-

ceptowalny: jak pokazuje tabela 3.60 hipotezy takiej nie można odrzucić na poziomie istotności 0,05.

Biorąc pod uwagę, że wszystkie trzy zmienne mają takie same kategorie można sformułować kolejne modele. Przykładowo, możliwa jest modyfikacja modelu przez dodatkowe uwzględnienie specyfiki kategorii, opisujących sytuację, gdy respondent relatywnie częściej (rzadziej) wskazywał na tę samą odpowiedź w każdym badaniu niż wynikałoby to z modelu $[XY][XZ][YZ]$. Innymi słowy, prawdopodobieństwa $\pi_{111}^{XYZ}$, $\pi_{222}^{XYZ}$, $\pi_{333}^{XYZ}$, będą posiadały własne specyficzne parametry $q_{111}$, $q_{222}$, $q_{333}$. Model taki — oznaczony w tabeli 3.60 jako $Q[XY][XZ][YZ]$ —jest akceptowalny, przy czym jeśli porównamy go z modelem prostszym $[XY][XZ][YZ]$ za pomocą testu warunkowego okazuje się, że uwzględnienie trzech parametrów opisujących interakcję trzeciego rzędu nie jest istotne statystycznie: $G^2 = 5,80 - 4,38 = 1,32$, $df = 3$, $(p = 0,70)$.

Ze względu na to, że mamy do czynienia ze zmiennymi porządkowymi, model $[XY][XZ][YZ]$ można uprościć, tj. opisać związek pomiędzy wybraną parą zmiennych za pomocą hipotezy mniej złożonej, przykładowo hipotezy o jednakowej interakcji, ustalonego dystansu, itd. Model, który w tabeli 3.60 oznaczono $[XY_{UA}][XZ_{UA}][YZ_{UA}]$ głosi, że związek pomiędzy każdą parą zmiennych daje się opisać za pomocą hipotezy o jednakowej interakcji[38]. Oznacza to, że w każdej podzbiorowości wyróżnionej ze względu na zmienną $Z$ wszystkie lokalne stosunki szans dla zmiennych $X$ i $Y$ są sobie równe i podobnie jest w odniesieniu do związku między parą zmiennych $X$ i $Z$ względem $Y$ oraz parą $Y$ i $Z$ względem $X$. Jak pokazuje tabela 3.60 model ten nie może być zaakceptowany na standardowo przyjmowanych poziomach istotności.

Model powyższy można zmodyfikować uwzględniając specyfikę głównej przekątnej, dla każdej pary zmiennych. Przykładowo, oznaczałoby to, że relatywnie częściej (rzadziej) współwystępują te same odpowiedzi na pytania zadane w 2000 i 2003 roku w podzbiorowościach wyróżnionych ze względu na odpowiedzi udzielane w 2005 roku. Tendencja ta jest modelowana za pomocą jednego parametru $q^{XY}$, w tym sensie główna przekątna uwzględniona jest podobnie jak w modelu $Q_hN$ a nie QN[39] (porównaj tabele 3.16 oraz 3.12). Podobnie uwzględniamy współwystępowanie tych samych odpowiedzi w 2000 i 2005 jak również 2003 i 2005 roku. Tak sformułowany model został oznaczony $[XY_{QhUA}][XZ_{QhUA}][YZ_{QhUA}]$. Jego dopasowanie — jak pokazują wyniki weryfikacji — jest bardzo dobre.

[38] Model o jednakowej interakcji dla trzech par zmiennych został szczegółowo opisany w rozdziale drugim, formalnie zdefiniowany za pomocą formuły 2.65.

[39] Gdy mamy do czynienia z trzema kategoriami każdej zmiennej uwzględnienie parametru specyficznego dla każdej komórki nie było możliwe, gdyż model taki wykorzystywałby tyle samo parametrów co model $[XY][XZ][YZ]$ a zakładałby dodatkowo quasi–symetrię.

Dodatkowo możliwe jest uwzględnienie — podobnie jak w opisanym powyżej modelu $Q[XY][XZ][YZ]$ — częstszego współwystępowania tej samej kategorii trzech zmiennych, tj. udzielanie tej samej odpowiedzi we wszystkich trzech punktach czasowych. Test warunkowy porównujący modele $[XY_{QhUA}][XZ_{QhUA}][YZ_{QhUA}]$ oraz $Q[XY_{QhUA}][XZ_{QhUA}][YZ_{QhUA}]$ pokazuje, że modyfikacja tego rodzaju nie poprawia dopasowania modelu w sposób istotny statystycznie: $G^2 = 3,28$, $df = 3$, $p = 0,3492$.

Powyższe analizy wskazują na model $[XY_{QhUA}][XZ_{QhUA}][YZ_{QhUA}]$. Można zadać pytanie czy nie można go uprościć pomijając parametr przekątnej dla wybranej pary zmiennych. Trzy tego typu modele zostały przedstawione w trzech ostatnich wierszach tabeli 3.60. Tylko model $[XY_{UA}][XZ_{QhUA}][YZ_{QhUA}]$ jest akceptowalny na poziomie istotności 0,01. Porównanie modeli $[XY_{UA}][XZ_{QhUA}][YZ_{QhUA}]$ oraz $[XY_{QhUA}][XZ_{QhUA}][YZ_{QhUA}]$ za pomocą testu warunkowego wskazuje ($G^2 = 3,63$, $df = 1$, $p = 0,0566$), że nie jest konieczne uwzględnienie parametrów $q^{XY}$.

Model prostszy stanowi więc adekwatny opis analizowanych danych. Parametry jednakowej interakcji wynoszą odpowiednio $\delta^{XY} = 2,17$, $\delta^{YZ} = 2,04$, $\delta^{XZ} = 1,39$. Przypomnijmy, że opisują one lokalne stosunki szans. Przykładowo wielkość $\delta^{YZ}$ wskazuje, że w grupie osób jednorodnej, ze względu na to, jakiej odpowiedzi udzieliły w 2000 roku, proporcja liczby osób, którym w 2005 roku problemy finansowe zdarzały się „często" do liczby osób, które udzieliły odpowiedzi „zdarzyły się" jest ponad dwukrotnie większa wśród osób które w 2003 roku udzieliły odpowiedzi „zdarzyły się" do liczby osób które udzieliły odpowiedzi „rzadko". Podobnie można zinterpretować pozostałe parametry. Zwraca uwagę, że relatywnie najniższa jest wartość $\delta^{XZ}$, co wydaje się o tyle sensowne, że zmienne te dotyczą dwóch badań, które dzieliła najdłuższa przerwa, co mogło przełożyć się na najsłabszy związek między tymi zmiennymi. Opisane powyżej zależności powinny jednak być zmodyfikowane, przez uwzględnienie, że porównując badania w 2003 i 2005 roku jak również 2000 i 2005, respondenci relatywnie częściej wskazywali na te same odpowiedzi niż wynikałoby z modelu jednakowej interakcji. Odpowiednie parametry wynoszą $q^{YZ} = 1,36$ oraz $q^{XZ} = 1,27$.

### 3.4.3 Zmiany w rozkładzie łącznym zmiennych — dane panelowe

Dane panelowe umożliwiają porównywanie czy dla badanych osób (obiektów) zmienia się bądź też pozostaje identyczna charakterystyka wynikająca z rozkładu łącznego dwóch lub większej liczby zmiennych. Przykład takich danych — pochodzących z drugiej i czwartej edycji „Diagnozy Społecznej" — pokazuje tabela 3.61. Kolejne wiersze opisują kombinacje dwóch zmiennych opisujących cechy respondenta w 2003 roku.

Zmienna $X_1$ wskazuje, czy respondenci byli w momencie badania bezrobotni czy też nie. Zmienna $Y_1$ opisuje opinie dotyczące tego, czy zmiany, jakie zaszły w Polsce po 1989 roku miały wpływ na życie respondenta, a jeśli tak, to czy ten wpływ był korzystny czy też niekorzystny.

Pytanie to było już wykorzystane w poprzednich analizach, opis operacjonalizacji tej zmiennej został zamieszczony przy okazji omówienia tabeli 3.50. Dla potrzeb analizy połączone zostały kategorie[40] *bardzo niekorzystny* i *raczej niekorzystny*, podobnie kategorie *bardzo korzystny* i *raczej korzystny*. Kolumny wskazują na kombinacje analogicznych zmiennych — oznaczonych jako $X_2$ oraz $Y_2$ - opisujących sytuację respondenta w 2007 roku. Tak więc, pierwsza kolumna opisuje osoby, które w 2007 roku były bezrobotne i wskazywały, że zmiany po 1989 roku były dla nich niekorzystne.

Tabela 3.61: Ocena zmian w 1989 roku przez osoby bezrobotne i pozostałe w 2003 (zmienne $X_1$ i $Y_1$) i 2007 roku[a] (zmienne $X_2$ i $Y_2$)

| | Wpływ zmian | Bezrobotni($X_2=1$) | | | Pozostali($X_2=2$) | | |
|---|---|---|---|---|---|---|---|
| | po 1989 roku: | $Y_2=1$ | $Y_2=2$ | $Y_2=3$ | $Y_2=1$ | $Y_2=2$ | $Y_2=3$ |
| Bezrobotni | Niekorzystny ($Y_1=1$) | 54,7 | 14,5 | 7,8 | 87,6 | 70,4 | 32,5 |
| ($X_1=1$) | Brak wpływu ($Y_1=2$) | 13,2 | 25,2 | 3,7 | 17,1 | 77,1 | 17,3 |
| | Korzystny ($Y_1=3$) | 4,2 | 3,1 | 1,0 | 5,1 | 11,0 | 11,0 |
| Pozostali | Niekorzystny ($Y_1=1$) | 39,8 | 22,6 | 9,2 | 619,1 | 562,1 | 234,9 |
| ($X_1=2$) | Brak wpływu ($Y_1=2$) | 13,0 | 47,9 | 10,7 | 261,7 | 797,7 | 197,3 |
| | Korzystny ($Y_1=3$) | 0,8 | 2,3 | 4,9 | 68,6 | 145,2 | 239,2 |

[a]Źródło: Diagnoza Społeczna 2003, 2007.

Metodę analizy danych tego typu zaproponował Duncan[41] (1980, 1981). Często stosowanym modelem do analizy danych tego typu jest symetria. Zauważmy, że na głównej przekątnej tej tabeli występują osoby, dla których kombinacja obydwu zmiennych jest identyczna w obydwu latach. Hipoteza symetrii odnosi się do komórek położonych poza nią. Zgodnie z nią zachodzi:

$$\pi_{ijkl}^{X_1Y_1X_2Y_2} = \pi_{klij}^{X_1Y_1X_2Y_2} \tag{3.63}$$

Przykładowo, oznacza to, że $\pi_{2113}^{X_1Y_1X_2Y_2} = \pi_{1321}^{X_1Y_1X_2Y_2}$. Pierwsze z prawdopodobieństw wskazuje na częstość osób, które w 2003 roku nie były bezrobotne a w 2007 roku ich

[40]Pomimo, połączenia kategorii w tablicy rozkładu łącznego kilka komórek ma niewielkie liczebności, co jest typową sytuacją, gdy rozpatrujemy rozkład wielu zmiennych. Niemniej dla modeli, które zamieszczamy w tej sekcji spełnione są warunki dotyczące liczebności podane w rozdziale pierwszym.

[41]Zagadnienia te omówione są również w (Fingleton 1984).

Tabela 3.62: Wyniki weryfikacji modeli dla danych z tabeli 3.61

| Model | df | $\chi^2$ | $G^2$ | $\Delta$ | BIC |
|---|---|---|---|---|---|
| $S$ | 15 | 319,9 ($p < 0,0001$) | 342,3 ($p < 0,0001$) | 9,7 | 218,9 |
| $SX_1Y_1X_2Y_2$ | 12 | 13,6 ($p = 0,3246$) | 15,0 ($p = 0,2396$) | 1,1 | -83,7 |
| $QS$ | 10 | 13,6 ($p = 0,1908$) | 14,9 ($p = 0,1364$) | 1,1 | -67,4 |
| $QS[X_1Y_1]_{UA}[X_2Y_2]_{UA}$ | 12 | 16,0 ($p = 0,1919$) | 17,2 ($p = 0,1408$) | 1,2 | -81,5 |
| $QS\{[X_1Y_1]_{UA} = [X_2Y_2]_{UA}\}$ | 13 | 16,0 ($p = 0,2508$) | 17,4 ($p = 0,1837$) | 1,2 | -89,6 |

status się zmienił, a jednocześnie ich opinia co do wpływu wydarzeń po 1989 roku na ich życie uległa zmianie: w 2003 roku odpowiedziały, że wpływ ten był niekorzystny a cztery lata później, że korzystny. Zgodnie z tym modelem jest ona równa częstości osób, dla których zmiana nastąpiła w odwrotnym kierunku dla obydwu cech, tj. nie były bezrobotne w 2003 a były w 2007 roku, a ich ocena zmian z 1989 roku zmieniła się z kategorii „korzystny" na „niekorzystny". Model taki — jak pokazuje tablica 3.62 — jest nierealistyczny. Podobnie jak w przypadku dwóch zmiennych słabe dopasowanie tego modelu wynika, z tego, że pełna symetria jest możliwa tylko wówczas, gdy rozkłady brzegowe takiej tabeli są identyczne, tj. rozkłady łączne obydwu zmiennych są identyczne w obydwu porównywanych latach. Mówiąc inaczej rozkład łączny zmiennych $X_1$ i $Y_1$ jest taki sam jak zmiennych $X_2$ i $Y_2$.

Bardziej realistyczny jest model quasi–symetrii, który zakładałby, że symetryczne są jedynie stosunki szans w tak skonstruowanej tabeli. Jak pokazuje tabela 3.62, dopasowanie takiego modelu jest bardzo dobre, mógłby on być zaakceptowany na poziomie istotności 0,05, jedynie nieco ponad 1% osób zaklasyfikowanych jest niezgodnie z tym modelem. Obydwa modele — S i QS — różni założenie o identyczności rozkładów brzegowych. Test warunkowy porównujący obydwa modele można potraktować jako weryfikację hipotezy głoszącej, że rozkład łączny obydwu zmiennych, pierwszej zdającej sprawę z sytuacji na rynku pracy ($X$) i opiniami na temat wpływu zmian po 1989 roku ($Y$) był taki sam w obydwu porównywanych badaniach. Wyniki testu warunkowego pokazują, ze założenie to należy odrzucić: ($G^2 = 342,3 - 14,9 = 327,4$, $df = 5$, $p < 0,0001$).

Zauważmy, że testowaną powyżej hipotezę o braku zmian w rozkładzie łącznym zmiennych $X$ oraz $Y$, można przedstawić jako połączenie następujących warunków:

1. Rozkład zmiennej $X_1$ jest taki sam jak rozkład zmiennej $X_2$, tj. odsetek osób bezrobotnych jest w obydwu badaniach identyczny.

2. Rozkład zmiennej $Y_1$ jest taki sam jak rozkład zmiennej $Y_2$, tj. w jednym i w drugim badaniu taki sam odsetek badanych wskazywał, że zmiany po 1989 roku miały wpływ korzystny, podobnie taki sam odsetek wskazywał na wpływ niekorzystny bądź brak wpływu.

3. Związek pomiędzy $X_1$ i $Y_1$ mierzony stosunkami szans jest taki sam jak pomiędzy $X_2$ i $Y_2$. Oznacza to, że nie zmieniły się wzór i siła związku pomiędzy „byciem bezrobotnym" a opiniami o zmianach jakie zaszły po 1989 roku.

Przypuśćmy, że jesteśmy zainteresowani przetestowaniem jedynie trzeciego z wymienionych warunków a nie wszystkich trzech łącznie. W tym celu do sformułowanej poprzednio hipotezy o quasi–symetrii, należy dodać warunek głoszący, że stosunki szans pomiędzy $X_1$ i $Y_1$ są analogiczne jak stosunki szans pomiędzy $X_2$ i $Y_2$. Wiążę się on z $(r-1)(c-1)$ dodatkowymi założeniami, to znaczy tyle lokalnych stosunków szans można wyznaczyć, gdy zmienna $X_1$ ma $r$ kategorii a liczba kategorii zmiennej $Y_1$ wynosi $c$. Model ten został oznaczony[42] jako $SX_1Y_1X_2Y_2$. Model ten jest akceptowalny. Porównanie tego modelu za pomocą testu warunkowego z modelem QS stanowi weryfikację interesującego nas założenia o identyczności związku w 2003 i 2007 roku. Okazuje się, że dopasowanie obydwu hipotez jest niemal identyczne, hipotezy powyższej nie można odrzucić ($G^2 = 15,03 - 14,88 = 0,15$, $df = 2$, $p < 0,9263$). Różnice w związku pomiędzy zmiennymi okazały nieistotne statystycznie dla obydwu badań.

Warto zauważyć, że powyższy test warunkowy, nie jest tym samym, co hipoteza $[ZX][ZY][XY]$, gdzie $Z$ wskazuje na rok badania. Podobnie, testowanej wcześniej hipotezy o identyczności rozkładu łącznych zmiennych $X_1$ i $Y_1$ oraz rozkładu $X_2$ i $Y_2$ nie możemy sprowadzić do hipotezy $[Z][XY]$. Danych panelowych nie można traktować tak, jakby pochodziły z dwóch niezależnych od siebie badań, przeprowadzonych w kolejnych latach, nie mamy bowiem do czynienia z niezależnymi od siebie próbami. Przeciwnie, w obydwu punktach czasowych rozpatrujemy te same jednostki.

W modelach przedstawionych do tej pory nie rozważaliśmy możliwości uwzględnienia porządkowego charakteru zmiennych $Y_1$, $Y_2$ opisujących opinie na temat wpływu zmian po 1989 roku. Można na przykład uprościć model QS zakładając, że związek pomiędzy nimi opisywany jest za pomocą modelu jednakowej interakcji. Model taki oznaczony został jako $QS[X_1Y_1]_{UA}[X_2Y_2]_{UA}$. Zarówno w 2000 jak i 2003 roku interakcję można opisać za pomocą jednego a nie dwóch (tj. $(r-1)(c-1)$) parametrów jak w modelu QS. W konsekwencji model ten będzie miał o dwa stopnie swobody więcej. Jak widać z tabeli 3.62 jest to model akceptowalny i test warunkowy porównujący

[42] Oznaczenie takie przyjęliśmy za Fingletonem (1984). Wynika ono stąd, ze stosunku do modelu symetrii rozkłady zmiennych $X_1$, $Y_1$, $X_2$, $Y_2$ muszą odzwierciedlać dane z próby, jeśli szacujemy rozkład oczekiwany metodą największej wiarygodności.

obydwa modele pokazuje, że uproszczenie modelu nie pogarsza jego dopasowania w sposób istotny statystycznie: ($G^2 = 17,2 - 14,9 = 0,31$, $df = 2$, $p = 0,5769$).

Parametry interakcji wynoszą odpowiednio: $\delta_1 = 1,55$ dla 2003 roku i $\delta_2 = 1,48$ dla 2007 roku. Oznaczałoby to że siła zależności jest słabsza dla roku 2007 (parametr jest bliższy wartości 1, czyli niezależności). Hipotezę można dodatkowo uprościć, zakładając, że siła zależności jest taka sama dla obydwu lat, tj. $\delta_1 = \delta_2$. Zauważmy, że model taki — oznaczony jako $QS\{[X_1Y_1]_{UA} = [X_2Y_2]_{UA}\}$ — jest szczególnym przypadkiem modelu $QS[X_1Y_1]_{UA}[X_2Y_2]_{UA}$ (jak również rozważanego wcześniej modelu $SX_1Y_1X_2Y_2$). Okazuje się, że model taki jest akceptowalny a test warunkowy porównujący modele $QS[X_1Y_1]_{UA}[X_2Y_2]_{UA}$ oraz $QS\{[X_1Y_1]_{UA} = [X_2Y_2]_{UA}\}$ pokazuje, że założenia o identycznej sile zależności opisywanej przez jednakową interakcję nie można odrzucić na standardowo przyjmowanych poziomach istotności: $G^2 = 17,35 - 17,24 = 0,11$, $df = 1$, $p = 0,94$. Parametr siły związku dla tego modelu wynosi: $\delta = 1,5$. Oznacza to — przykładowo — że proporcja liczby osób które uważają, że wpływ zmian po 1989 roku był korzystny do liczby, które nie dostrzegają wpływu jest 1,5 razy mniejsza wśród bezrobotnych niż pozostałych osób.

Jak zostało zasygnalizowane wcześniej, sformułowane powyżej modele mogą być wykorzystane do analizy danych inne niż panelowe. Podobną strukturę miałyby tablice, w których zestawiłoby się dwie cechy respondentów i ich ojców, przykładowo wykształcenia i zawodu. W tym sensie powyższe modele mogą być użyteczne również do opisu ruchliwości społecznej. Podobnie możliwe byłoby analizowanie wzorów zawierania małżeństw pod kątem kilku cech mężów i żon. Warto podkreślić, że w rozdziale tym przytoczone zostały jedynie wybrane modele. Można przykładowo konstruować modele pośrednie pomiędzy modelem $S$ i modelem $SX_1Y_1X_2Y_2$. Ponadto, do opisu związku między zmiennymi można wykorzystać inne hipotezy niż jednakowa interakcja.

# Podsumowanie

W pracy zaprezentowane zostały możliwości analizowania rozkładu dwóch lub większej liczby zmiennych za pomocą modeli logarytmiczno–liniowych. Jak zostało pokazane w rozdziale pierwszym, w ramach tej metody sformułować można wiele hipotez określających związek pomiędzy zmiennymi. Możliwości te znacznie się poszerzają, jeśli uwzględni się porządkowy charakter jeden bądź kilku zmiennych, co zostało zaprezentowane w rozdziale drugim. Dodatkowe modele można formułować dla tablic, gdzie kategorie dwóch (lub większej liczby zmiennych) są identyczne, jak dzieje się na przykład w odniesieniu do danych panelowych lub tablic ruchliwości, czemu poświęcony został rozdział trzeci.

Praca wykorzystuje nieco inny schemat prezentacji poszczególnych modeli, aniżeli można zazwyczaj spotkać w literaturze, gdzie wychodzi się od postaci parametrycznej modelu. Prezentację modelu rozpoczyna sformułowanie hipotezy, przy wykorzystaniu pojęć równomierności rozkładu, niezależności stochastycznej, stosunków szans. Niektóre modele wskazują również na inne specyficzne relacje pomiędzy prawdopodobieństwami (model quasi–niezależności, model dystansu, modelowanie asymetrii). Ta sama hipoteza, może być parametryzowana na wiele różnych sposobów a wybór odpowiedniej parametryzacji ma pomóc w interpretacji wybranych aspektów związanych z opisem rozkładów poszczególnych zmiennych, bądź związku między zmiennymi.

Oczywiście model w postaci parametrycznej również może być wykorzystywany do formułowania poszczególnych hipotez. Jak zostało pokazane w rozdziale pierwszym, polega to na nakładaniu na model nasycony określonych warunków. Trzeba jednak wówczas pamiętać z jakiego rodzaju parametryzacją mamy do czynienia. Przykładowo, definicja quasi–niezależności jako modelu, w którym każda komórka na głównej przekątnej posiada własny parametr, może być myląca, gdyż ten sam model można formułować również w inny sposób, co pokazuje tabela 3.15. Aby powyższa formula była jednoznaczna konieczne jest jej doprecyzowanie o definicje poszczególnych parametrów modelu. Warto też zwrócić uwagę, że trudności w interpretacji parametrów modelu są jeszcze większe, gdy wykorzystuje się jego addytywną postać — najbar-

dziej rozpowszechnioną w literaturze — gdyż ich wartości są odnoszone do logarytmu prawdopodobieństwa.

Związek między zmiennymi jest w modelach logarytmiczno–liniowych opisywany przede wszystkim przez stosunki szans i w tym sensie formułowanie hipotez za ich pomocą wydaje się bardziej precyzyjne. Ten typ prezentacji pozwala dostrzec różnice pomiędzy modelami bądź ich równoważność w niektórych sytuacjach (przykładowo, modelu quasi–niezależności i quasi–symetrii dla tablicy o wymiarach 3 x 3, modeli QS i QDS jak również, QCP i QFD dla tablic o wymiarach 4 x 4). Niektóre formuły oparte na stosunkach szans są dość skomplikowane i w tym sensie dla lepszego zrozumienia hipotezy dobrze jest prezentować model wykorzystując zarówno stosunki szans (bądź inne relacje pomiędzy prawdopodobieństwami) jak też sformułowanie parametryczne.

Warto jeszcze raz podkreślić, że wybór określonego modelu, nie powinien opierać się jedynie na przesłankach statystycznych, tj. dopasowaniu modelu do danych. Istotna wydaje się teoria badanego zjawiska jak również przypuszczenia badacza dotyczące wzoru związku pomiędzy zmiennymi. Oczywiście, przesłanki te nie zawsze prowadzą do wskazania jednego konkretnego modelu, na ogół teoria w naukach społecznych nie jest do tego stopnia precyzyjna. Niemniej teoria badanego zjawiska może na przykład sugerować, które modele powinno się ze sobą porównywać. W pracy powyższej analizując konkretne dane empiryczne, przesłanki tego typu były brane pod uwagę w ograniczonym zakresie. Wynika to jednak ze specyfiki pracy, która jest nastawiona na prezentację samych modeli, ich własności oraz możliwości ich zastosowania. Przykłady — jak zostało zaznaczone na wstępie — miały charakter ilustracyjny.

Warto zasygnalizować kwestie, które wykraczały poza ramy tej pracy, ale są ściśle związane z omawianą metodą. Często do analizy związku pomiędzy zmiennymi wykorzystuje się, tzw. *modele logitowe* (*logit models*, *effect models*) (Goodman 1973, Agresti 1984, Aldrich i Nelson 1984, Hagenaars 1990, Liao 1994, Demaris 1998, Borooah 2001). W tym ujęciu jedną ze zmiennych traktuje się — podobnie jak w równaniu regresji liniowej — jako zmienna zależną. Dokładniej jest nią logit (tj. logarytm szansy) dla dwóch wyróżnionych kategorii tej zmiennej. Stosunkowo łatwo przekształcić model logarytmiczno–liniowy w odpowiedni model logitowy, obydwie metody są ze sobą silnie powiązane. Choć w tej pracy żaden model logitowy nie został sformułowany, interpretacja parametrów często była bliska temu ujęciu. Podejście to często rozszerza się, formułuje się modele analogiczne do analizy ścieżkowej, co pozwala uwzględnić bardziej skomplikowane zależności przyczynowo-skutkowe. Warto również wspomnieć o regresji logistycznej, która jest pod wieloma względami podobna do modelu logitowego, przy czym można w niej uwzględnić dodatkowo zmienne mierzone na skali interwałowej (Dobson 2002, Pampel 2006, O'Connell 2006).

Inną kwestią nie omówioną w pracy jest wykorzystanie danych panelowych do modelowania związków przyczynowo–skutkowych (Duncan 1975, 1981, Fingleton 1984, Karpiński 2006) . W modelach takich wykorzystuje się założenie, że zmienna $X$ mierzona punkcie czasowym $t_1$ na ogół nie może zależeć od cechy $Y$ mierzonej w późniejszym punkcie $t_2$. Rozpatrując kilka zmiennych jednocześnie, można próbować ustalać związki przyczynowo–skutkowe między zmiennymi $X$ i $Y$.

Istnieje również możliwość uwzględnienia w modelach logarytmiczno–liniowych zmiennych *ukrytych* (*latent variables*) (Lazarsfeld 1950, Lazarsfeld i Henry 1968, Goodman 1974, McCutcheon 1987, Hagenaars 1990, Hagenaars 1993, Clogg 1995, Nawojczyk i McCutcheon 1996, Hagenaars i McCutcheon 2002). Zmienne obserwowalne traktuje się wówczas jako niedoskonałe wskaźniki interesującej nas cechy, której nie można mierzyć bezpośrednio. W tym sensie daje się zauważyć pewną analogię tego typu modeli do analizy czynnikowej i modeli strukturalnych, które jednak na ogół wymagają zmiennych zmierzonych na skali interwałowej. Modele ze zmiennymi ukrytymi stosuje się również w odniesieniu do tablic ruchliwości (Boudon 1973, Clogg 1981, Duncan 1985) jak też danych panelowych (Markus 1984, Tomaszewski 2004).

Na koniec warto zasygnalizować, jakie oprogramowanie wykorzystuje się do estymacji modeli logarytmiczno–liniowych. Wiele z nich daje się analizować za pomocą najbardziej popularnych pakietów, takich jak SPSS, SAS, STATISTICA, STATA. Duże możliwości daje program R i powiązany z nim pakiet S–PLUS. Obliczenia wykorzystane w tej pracy, zostały wykonane za pomocą pakietu LEM, przeznaczanego głownie do analizy zmiennych jakościowych (nominalnych i porządkowych). Jego autorem jest Jeroen Vermunt (1997). Program ten pozwala również estymować modele z uwzględnieniem zmiennych ukrytych, choć pod tym względem większe możliwości dają pakiety LATENT GOLD i MPLUS.

# Aneks A

# Opis zbiorów danych, dodatkowe ilustracje i formuły

Aneks składa się z kilku części. W pierwszej z nich zamieszczone zostały informacje o zbiorach danych wykorzystanych w tej pracy. W drugiej znajduje się tzw. twierdzenie o agregacji w wersji ogólnej, które zostało zasygnalizowane w rozdziale pierwszym. Następnie, przedstawiony został przykład, który ilustruje niejednoznaczność dekompozycji modelu zgodnego z hipotezą o quasi–niezależności na podzbiorowości wyróżnione ze względu na zmienną $Z$, przy zdefiniowaniu modelu przez formuły 3.4, 3.5, 3.8. W dalszej części przedstawione są wybrane modele uwzględniające specyfikę głównej przekątnej sformułowane za pomocą lokalnych stosunków szans, wraz z uzasadnieniem liczby stopni swobody dla tych modeli. Na końcu pokazana zostanie równoważność modeli QS i QDS jak również modeli QCP i QFD dla tablic o wymiarach 4 x 4.

## A.1 Informacje o zbiorach danych wykorzystanych do przykładów empirycznych

W pracy zostały do ilustracji empirycznych wykorzystane następujące zbiory danych:

- **Europejski Sondaż Społeczny**, Paweł. B. Sztabiński, Franciszek Sztabiński, Ośrodek Realizacji Badań Socjologicznych, Instytut Filozofii i Socjologii Polskiej Akademii Nauk, Warszawa. Projekt badawczy finansowany ze środków Komitetu Badań Naukowych. Przeprowadzono cztery tury tego badania. Począwszy od 2002 roku badanie powtarzane jest co dwa lata. W tabelach 1.27, 2.24, 3.8, 3.9 wykorzystano dane dla pierwszej edycji z 2002 roku. W badaniu tym odsetek zrealizowanych wywiadów wyniósł 73,2%, przeprowadzo-

no 2110 wywiadów. Dane z tabeli 2.9 dotyczyły II tury ESS przeprowadzonej w 2004 roku. Analogiczny odsetek zrealizowanych wywiadów wyniósł 70,2 %, przeprowadzono 1721 wywiadów. Jeśli chodzi o dane dla tabel 2.27, 3.55 to pochodzą one z I tury badania, przy czym obejmują również dane z innych krajów. Metodologia badania była pod wieloma względami podobna, niemniej kraje różniły się ze względu na schemat doboru próby jak również odsetek zrealizowanych wywiadów. Więcej informacji o badaniu można znaleźć na stronie *www.europeansocialsurvey.org*.

- **Polskie Generalne Sondaże Społeczne**, skumulowany komputerowy zbiór danych 1992-2008. Bogdan Cichomski (kierownik programu), Tomasz Jerzyński, i Marcin Zieliński. Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa 2009. Badanie jest realizowane od 1992 roku i było powtarzane dziewięciokrotnie. W tabeli 2.20 wykorzystane zostały dane z 1997 roku, wówczas przebadano 75,1 % założonej próby, co przekłada się na 2402 wywiady. Dane z 2005 roku kiedy analogiczny odsetek wynosił 60,6% (1277 osób) zostały wykorzystane w tabelach 2.5 oraz 2.19. Więcej informacji o doborze próby, konstrukcji wagi można znaleźć na stronie *pgss.iss.uw.edu.pl.*

- **Diagnoza Społeczna, Warunki i Jakość Życia Polaków,**, Rada Monitoringu Społecznego, Warszawa 2007. Badanie ma charakter panelowy, pierwsza edycja badania miała miejsce w 2000 roku, kolejne w 2003, 2005, 2007 i 2009 roku. Badanie z jednej strony dotyczy gospodarstw domowych, z drugiej strony indywidualnych ankieterów, tj. istnieją dwie wersje ankiety. Tabela 3.6 dotyczy gospodarstw domowych przebadanych w 2000 i 2005 roku. Pozostałe tabele dotyczą badania skierowanego do indywidualnych respondentów. Obejmują one odpowiednio lata: tabela 3.50 — 2003 i 2005 rok, tabela 3.59 — 2000, 2003 i 2005 rok, tabela 3.61 — 2003 i 2007 rok. Kwestia odsetka realizacji próby jest bardziej skomplikowana ze względu na panelowy charakter badania. Więcej informacji o badaniu, doborze próby, zastępowaniu jednostek, które nie wzięły udziału w badaniu, w kolejnych turach panelu, konstrukcji wagi można znaleźć w: Czapliński i Panek (2007) oraz na stronie *www.diagnoza.com.*

- **Struktura Społeczna II**, badanie zostało przeprowadzone w 1987 roku w Instytucie Filozofii i Socjologii PAN, Kierownikiem badania był Kazimierz Słomczyński, w skład zespołu wchodzili również Ireneusz Białecki, Henryk Domański, Krystyna Janicka, Bogdan Mach, Zbigniew Sawiński, Joanna Sikorska i Wojciech Zaborowski. Badanie objęło osoby wieku 21-65 lat. Więcej informacji na

temat badania można znaleźć w książce "Struktura Społeczna: schemat teoretyczny i warsztat badawczy" (1989).

## A.2 Twierdzenie o agregacji (collapsibility theorem)

Twierdzenie poniższe opisuje w jakich sytuacjach warunkowe stosunki szans wyróżnione dla dwóch zmiennych względem trzeciej zmiennej są równe stosunkom szans wyznaczonych dla tych samych kategorii w całej zbiorowości.

**Twierdzenie o agregacji, (Bishop i inni, 1975)**
Zmienne będące przedmiotem zainteresowania, dzielimy na trzy grupy:

$$X_1, X_2, X_3, \ldots$$

$$Y_1, Y_2, Y_3, \ldots$$

$$Z_1, Z_2, Z_3, \ldots$$

Przez $Z$ oznaczmy zmienną złożoną uwzględniającą kombinacje wszystkich zmiennych $Z_1, Z_2, Z_3, \ldots$. Przyjmijmy, że każda ze zmiennych $X$ jest warunkowo niezależna stochastycznie od zmiennej $Z$ względem pozostałych zmiennych np. $Y$. Rozpatrujemy rozkłady:

1. w podzbiorowościach wyróżnionych ze względu na wartości zmiennej $Z$
2. brzegowe powstałe przez zsumowanie po indeksach zmiennej $Z$

Stosunki szans uwzględniające *przynajmniej jedną ze zmiennych* $X$ dla określonych w ten sposób rozkładów warunkowych i brzegowych są sobie równe. *Dowód: porównaj Bishop i inni, 1975, str. 75.*

Dla przykładu, jeżeli badamy stosunki szans dla dowolnych kategorii zmiennych $X_1$ i zmiennych $Y_1$ to zachodzi

$$\Theta^{X_1;\,Y_1\,(Z)}_{a/b;c/d(k)} = \Theta^{X_1;\,Y_1}_{a/b;c/d}$$

Należy podkreślić, że równość ta zachodzi tylko wówczas jeśli jedną ze zmiennych, pomiędzy którymi zależność mierzymy jest zmienna $X$, tj zmienna niezależna od zmiennej $Z$. Nie musi więc zachodzić $\Theta^{Y_1;\,Y_2\,(Z)}_{a/b;c/d(k)} = \Theta^{Y_1;\,Y_2}_{a/b;c/d}$.

## A.3 Ilustracja do modelu quasi–niezależności

Poniżej zamieszczony został przykład ilustrujący niejednoznaczność dekompozycji rozkładu łącznego zmiennych $X$ i $Y$, ilustrującego hipotezę o quasi–niezależności na trzy podzbiorowości: osób niemobilnych ($Z = 1$) oraz dwóch podzbiorowości osób mobilnych, opisywanych przez niezależność zmiennych $X$ i $Y$: dla jednej z nich nie wykluczamy możliwości współwystępowania tej samej kategorii dwóch zmiennych ($Z = 2$), dla drugiej wykluczamy taką sytuację ($Z = 3$). Odsetki poszczególnych podzbiorowości są oznaczone jako $\beta_1$, $\beta_2$, $\beta_3$. Model jest opisywany przez formuły 3.4, 3.5, 3.8.

Tabela A.1: Rozkład łączny zmiennych $X$ i $Y$ zgodny z hipotezą QN (w procentach)

| $X\backslash Y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 13,44 | 3,63 | 2,77 | 1,91 |
| 2 | 5,55 | 7,58 | 2,77 | 1,91 |
| 3 | 8,32 | 5,45 | 4,58 | 2,87 |
| 4 | 8,32 | 5,45 | 4,16 | 21,29 |

W tabeli A.1 podany jest przykład rozkładu spełniającego warunek quasi– niezależności. W tabeli A.2 przedstawione zostały dwie możliwości dekompozycji tego rozkładu: przedstawiono między innymi informację o wielkości poszczególnych podzbiorowości. W pierwszej sytuacji trzy podzbiorowości mają niezerowe częstości, w drugiej podzbiorowość opisywana w pełni przez niezależność dwóch zmiennych ($Z = 2$) nie występuje. Pomimo tych różnic, obydwie sytuacje odzwierciedlają ten sam rozkład przedstawiony w tabeli A.1. Przykładowo:

$$\pi_{11} = 0,1344 = 0,37 \cdot 0,2703 + 0,43 \cdot 0,08 + 0,2 \cdot 0,00 = 46,9 \cdot 0,2866 + 53,1 \cdot 0.$$

Przykład ten pokazuje również, że różne mogą być odsetki dla poszczególnych kategorii w pozdbiorowości osób niemobilnych. Można też zaobserwować, że stosunki szans dla podzbiorowości $Z = 2$ są równe 1. Podobnie równe 1 są stosunki szans dla podzbiorowości $Z = 3$, wyznaczone dla kategorii leżących poza główną przekątną.

Tabela A.2: Rozkłady łączne zmienych $X$ i $Y$ w podzbiorowościach (dwóch lub trzech) wyróżnionych ze względu na zmienną $Z$, zgodne z rozkładem w tabeli A.1 (w procentach, wartości zaokrąglone)

| Trzy podzbiorowości | | | | | Dwie podzbiorowości, $\beta_2 = 0$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $Z = 1$, $\beta_1 = 37,0$ | | | | | $Z = 1$, $\beta_1 = 46,9$ | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | $X\backslash Y$ | 1 | 2 | 3 | 4 |
| 1 | 27,03 | 0,00 | 0,00 | 0,00 | 1 | 28,66 | 0,00 | 0,00 | 0,00 |
| 2 | 0,00 | 13,51 | 0,00 | 0,00 | 2 | 0,00 | 16,17 | 0,00 | 0,00 |
| 3 | 0,00 | 0,00 | 5,41 | 0,00 | 3 | 0,00 | 0,00 | 9,77 | 0,00 |
| 4 | 0,00 | 0,00 | 0,00 | 54,05 | 4 | 0,00 | 0,00 | 0,00 | 45,40 |
| $Z = 2$, $\beta_2 = 43,0$ | | | | | $Z = 2$, $\beta_2 = 0$ | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | | | | | |
| 1 | 8,00 | 6,00 | 4,00 | 2,00 | | | | | |
| 2 | 8,00 | 6,00 | 4,00 | 2,00 | | | | | |
| 3 | 12,00 | 9,00 | 6,00 | 3,00 | | | | | |
| 4 | 12,00 | 9,00 | 6,00 | 3,00 | | | | | |
| $Z = 3$, $\beta_3 = 20,0$ | | | | | $Z = 3$, $\beta_3 = 53,1$ | | | | |
| $X\backslash Y$ | 1 | 2 | 3 | 4 | $X\backslash Y$ | 1 | 2 | 3 | 4 |
| 1 | 0,00 | 5,26 | 5,26 | 5,26 | 1 | 0,00 | 6,84 | 5,22 | 3,60 |
| 2 | 10,53 | 0,00 | 5,26 | 5,26 | 2 | 10,44 | 0,00 | 5,22 | 3,60 |
| 3 | 15,79 | 7,89 | 0,00 | 7,89 | 3 | 15,66 | 10,26 | 0,00 | 5,40 |
| 4 | 15,79 | 7,89 | 7,89 | 0,00 | 4 | 15,66 | 10,26 | 7,83 | 0,00 |

## A.4 Dodatkowe formuły dla wybranych modeli

### A.4.1 Model quasi–niezależności QN

Hipotezę o quasi-niezależności można sformułować za pomocą lokalnych stosunków szans, jako połączenie trzech warunków:

$$\Theta^{XY}_{ab} = 1, \text{ gdzie } |\text{a-b}|>1, \tag{A.1}$$

$$\Theta^{XY}_{q(q+1)} = \Theta^{XY}_{(q+1)q} \text{ dla } 1 \leq q \leq r-2, \tag{A.2}$$

$$\Theta^{XY}_{qq} = \frac{1}{\Theta^{XY}_{(q-1)q}\Theta^{XY}_{(q+1)q}} \text{ dla } 2 \leq q \leq r-2. \tag{A.3}$$

Warunek A.1 określa, że wszystkie lokalne stosunki szans, nie zawierające komórek związanych z główną przekątną są równe 1. Przykładowo, dla tabeli 3.14 daje się na

przykład pokazać, że:

$$\Theta_{13}^{XY} = \frac{131,9 \cdot 64,2}{49,0 \cdot 172,7} = 1.$$

Podobnie równe 1 są stosunki szans $\Theta_{14}^{XY}$, $\Theta_{24}^{XY}$, $\Theta_{31}^{XY}$, $\Theta_{41}^{XY}$, $\Theta_{42}^{XY}$. Jak widać dla analizowanej tabeli takich warunków jest sześć. Jeśli chcielibyśmy ogólnie określić ich liczbę, należy zauważyć, że dotyczą one pseudo–przekątnych które, nie przylegają do głównej przekątnej, po obydwu jej stronach. Liczba takich stosunków szans wynosi[1]:

$$2[1 + \ldots + (r-3)] = 2(r-2)\frac{r-3}{2} = (r-2)(r-3).$$

Kolejny warunek A.2 dotyczy stosunków szans wyznaczonych dla komórek przylegających do głównej przekątnej, na przykład $\Theta_{12}^{XY}$, $\Theta_{23}^{XY}$ itd. Ponieważ stosunki szans tego typu obejmują komórki głównej przekątnej, wielkości te nie muszą być równe 1, natomiast w modelu quasi–niezależności są one symetryczne. Przykładowo, dla tabeli 3.14:

$$\Theta_{12}^{XY} = \frac{118,9 \cdot 172,7}{131,9 \cdot 134,5} = \Theta_{21}^{XY} = 1,16.$$

Podobnie, zachodzi:

$$\Theta_{23}^{XY} = \Theta_{32}^{XY};\ \Theta_{34}^{XY} = \Theta_{43}^{XY}.$$

Warunków takich jest $r-2$. Jeśli chodzi o ostatni warunek A.3 może się on wydawać na pierwszy rzut oka mało intuicyjny. Zauważmy jednak, że jest konsekwencją tego, że zgodnie z modelem quasi–niezależności warunkowe szanse, które nie obejmują komórek na głównej przekątnej są sobie równe, również wtedy, gdy jedna z nich leży powyżej głównej przekątnej, a druga z nich poniżej. Dla przykładu weźmy stosunek szans:

$$\Theta_{1/4;2/3}^{X\ \ Y} = \frac{\Omega_{1/4(2)}^{X\ (Y)}}{\Omega_{1/4(3)}^{X\ (Y)}} = \frac{\pi_{12}^{XY}/\pi_{42}^{XY}}{\pi_{13}^{XY}/\pi_{43}^{XY}} = 1.$$

Przypomnijmy, że dowolny stosunek szans może być przedstawiony jako iloczyn lokalnych stosunków szans (porównaj 2.1), dlatego analizowana powyżej wielkość jest równa:

$$\Theta_{1/4;2/3}^{X\ \ Y} = \Theta_{12}^{XY}\Theta_{22}^{XY}\Theta_{32}^{XY} = 1.$$

Przekłada się to na warunek A.3, tj.

$$\Theta_{22}^{XY} = \frac{1}{\Theta_{12}^{XY}\Theta_{32}^{XY}}.$$

Podobnie w omawianym przykładzie można pokazać, że:

$$\Theta_{33}^{XY} = \frac{1}{\Theta_{23}^{XY}\Theta_{34}^{XY}}.$$

[1]W tej i kilku kolejnych formułach dotyczących liczby stopni swobody będziemy posługiwać się formułą na sumę skończonego ciągu arytmetycznego, tj. $1 + 2 + \ldots + n = (n+1)(n/2)$.

Warunków tego typu jest $r-3$, jeśli mamy do czynienia ze zmiennymi o $r$ kategoriach.

Liczba niezależnych warunków A.1–A.3 określa liczbę stopni swobody modelu quasi–niezależności. Jest ich:

$$df = (r-2)(r-3)+(r-2)+(r-3) = (r-2)^2+(r-3) = r^2-3r+1 = (r-1)^2-r.$$

Model ten posiada $r$ stopni swobody więcej niż model niezależności, co oznacza, że posiada $r$ dodatkowych niezależnych parametrów. Tak więc w tabeli konieczne jest określenie każdego z parametrów $q_1, \ldots q_r$ leżących na głównej przekątnej.

### A.4.2 Model quasi–niezależności QhN

Hipoteza QhN głosi — podobnie jak zwykły model quasi–niezależności — warunek A.1. Warunki A.2 oraz A.3 wymagają przeformułowania, tak więc zgodnie z tą hipotezą zachodzi tj.

$$\Theta^{XY}_{ab} = 1, \text{ gdzie } |\text{a-b}|>1 \tag{A.4}$$

$$\Theta^{XY}_{j(j+1)} = \Theta^{XY}_{(j+1)j} = 1/q \text{ dla } 1 \leq j \leq r-2 \tag{A.5}$$

$$\Theta^{XY}_{jj} = q^2 \text{ dla } 1 \leq j \leq r-1. \tag{A.6}$$

Przypomnijmy, że ilustrację tego modelu stanowiła tabela 3.16. Warunki A.5 wskazują, że stosunki szans na pseudo–przekątnej przylegającej do głównej przekątnej są nie tylko symetryczne, ale również sobie równe. Warunków tych jest $2(r-2)-1$. Podobnie równe sobie są — zgodnie z formułą A.6 — stosunki szans na głównej przekątnej, co przekłada się na $(r-2)$ warunki. Należy jednak zauważyć, że stosunki szans z warunków A.2 oraz A.3 definiowane są przez tą samą wielkość $q$, co zmniejsza liczbę niezależnych warunków o 1. Łącznie formuły $A.4-A.6$ definiują $(r-1)^2-1$ niezależnych warunków, co określa liczbę stopni swobody tego modelu.

### A.4.3 Model QUA

Założenia dotyczące lokalnych stosunków szans, które definiują model QUA są następujące:

$$\Theta^{XY}_{ab} = \delta, \tag{A.7}$$

$$\Theta^{XY}_{q(q+1)} = \Theta^{XY}_{(q+1)q}, \tag{A.8}$$

$$\Theta^{XY}_{m(m-1)}\Theta^{XY}_{mm}\Theta^{XY}_{m(m+1)} = \left(\Theta^{XY}_{ab}\right)^3 = \delta^3, \tag{A.9}$$

gdzie $|a-b|>1$, jak również $1 \leq q \leq r-2$ oraz $2 \leq m \leq r-2$. Powyższe założenia modelu wymagają krótkiego omówienia. Przypuśćmy, że posługujemy się tablicą o wymiarach 5 x 5, do ich omówienia pomocna może być ilustracja modelu zamieszczona w tabeli 3.26. Warunek A.7 głosi — podobnie jak zwykły model jednakowej interakcji — równość lokalnych stosunków szans. Jednak w tym przypadku równość nie obejmuje komórek związanych z główną przekątną. Dla analizowanego przykładu oznacza to, że:

$$\Theta_{13}^{XY} = \Theta_{14}^{XY} = \Theta_{24}^{XY} = \Theta_{31}^{XY} = \Theta_{41}^{XY} = \Theta_{42}^{XY} = \delta$$

Zakładamy więc 5 równości. Bardziej ogólnie dla tablicy o wymiarach $r$ na $r$ ograniczeń tych jest $(r-3)(r-2)-1$. Warunek A.9 dotyczy symetrii stosunków szans wyznaczonych dla komórek przylegających do głównej przekątnej. W odniesieniu do analizowanego przykładu:

$$\Theta_{12}^{XY} = \Theta_{21}^{XY} = \delta/q_2$$

Podobnie można pokazać, że:

$$\Theta_{23}^{XY} = \Theta_{32}^{XY}, \quad \Theta_{34}^{XY} = \Theta_{43}^{XY}.$$

Dla analizowanej tabeli warunków tych jest trzy, ogólnie jest ich $r-2$. Ostatnie założenie na pierwszy rzut oka wydaje się najmniej intuicyjne. Pokazuje ono, że choć stosunki szans wyznaczone dla głównej przekątnej (np. $\Theta_{22}^{XY}$), bądź obejmujące komórki z nią związane (np. $\Theta_{12}^{XY}$) nie są — tak jak w oryginalnej postaci modelu jednakowej interakcji — równe wielkości stosunku szans określonego w warunku A.7, to jednak ta wielkość daje się przedstawić jako średnia geometryczna tych pierwszych, na przykład:

$$\Theta_{21}^{XY}\Theta_{22}^{XY}\Theta_{23}^{XY} = \Theta_{2/3;1/4}^{X\ \ Y} = \frac{\delta}{q_2}\delta q_2 q_3 \frac{\delta}{q_3} = \delta^3.$$

Powyższy zapis oznacza, że $\Theta_{2/3;1/4}^{X\ \ Y}$ – czyli stosunek szans porównujący pierwszą i czwartą kategorię zmiennej $Y$ dla drugiej i trzeciej kategorii zmiennej $X$ — jest równy wielkości $\delta$ — która opisuje lokalny stosunek szans, który nie obejmuje głównej przekątnej — podniesionej do potęgi trzeciej. Podobnie zachodzi:

$$\Theta_{32}^{XY}\Theta_{33}^{XY}\Theta_{34}^{XY} = \Theta_{3/4;2/5}^{X\ \ Y} = \frac{\delta}{q_3}\delta q_3 q_4 \frac{\delta}{q_4} = \delta^3$$

Ogólnie warunków tego typu jest $r-3$. Podsumowując ogólna liczba ograniczeń A.7–A.9 nakładanych na model, tj. liczba stopni swobody wynosi:

$$df = (r-3)(r-2) - 1 + (r-2) + (r-3) = r(r-3).$$

### A.4.4 Model Q(R=C)1

Zgodnie z hipotezą o quasi–symetrycznym modelu wierszowo-kolumnowym I typu z uwzględnieniem specyfiki głównej przekątnej $Q(R = C)1$ muszą być spełnione następujące warunki:

$$\Theta_{ij}^{XY} = \Theta_{ji}^{XY}, \tag{A.10}$$

$$\Theta_{i(i+k)}^{XY} = \delta_i \delta_{i+k}, \tag{A.11}$$

$$\frac{\left(\Theta_{qq}^{XY}/\Theta_{(q+1)(q)}^{XY}\right)\left(\Theta_{q(q-1)}^{XY}/\Theta_{(q+1)(q-1)}^{XY}\right)}{\left(\Theta_{q(q+1)}^{XY}/\Theta_{(q+1)(q+1)}^{XY}\right)\left(\Theta_{q(q+2)}^{XY}/\Theta_{(q+1)(q+2)}^{XY}\right)} =$$

$$= \left(\frac{\Theta_{(q+1)(q-1)}^{XY}/\Theta_{(q+2)(q-1)}^{XY}}{\Theta_{(q+1)q}^{XY}/\Theta_{(q+2)q}^{XY}}\right)^4 \tag{A.12}$$

$$\left(\frac{\Theta_{(m-2)m}^{XY}/\Theta_{(m-2)(m+1)}^{XY}}{\Theta_{(m-1)m}^{XY}/\Theta_{(m-1)(m+1)}^{XY}}\frac{\Theta_{(m-1)(m+1)}^{XY}/\Theta_{(m-1)(m+2)}^{XY}}{\Theta_{m(m+1)}^{XY}/\Theta_{m(m+2)}^{XY}}\right)^2 =$$

$$= \frac{\Theta_{mm}^{XY}/\Theta_{m(m+1)}^{XY}}{\Theta_{(m-1)m}^{XY}/\Theta_{(m-1)(m+1)}^{XY}} \tag{A.13}$$

dla $k \geq 3$, $2 \leq q \leq (r-3)$, $3 \leq m \leq (r-3)$, gdzie $r$ jest liczbą kategorii zmiennej wierszowej lub kolumnowej. Warunki te wymagają omówienia. Niektóre z nich mają zastosowanie, gdy mamy do czynienia z co najmniej sześcioma kategoriami zmiennej wierszowej i kolumnowej. Dlatego ilustracją będzie tabela A.3 i wyznaczone na jej podstawie lokalne stosunki szans zamieszczone zostały w tabeli A.4, przy czym w stosunku do tabeli A.3 zachodzi $\delta_1 = \psi_{i+1}/\psi_i$, np. $\delta_1 = \psi_2/\psi_1$, $\delta_2 = \psi_3/\psi_2$, itd.

Tabela A.3: Ilustracja modelu symetrycznego modelu wierszowo–kolumnowego I typu z uwzględnieniem specyfiki głównej przekątnej — parametry interakcji

| $X\backslash Y$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
|---|---|---|---|---|---|---|
| $x_1$ | $q_1$ | 1 | 1 | 1 | 1 | 1 |
| $x_2$ | 1 | $\psi_2^2 q_2$ | $\psi_2^2 \cdot \psi_3$ | $\psi_2^3 \cdot \psi_4$ | $\psi_2^4 \cdot \psi_5$ | $\psi_2^5 \cdot \psi_6$ |
| $x_3$ | 1 | $\psi_3 \cdot \psi_2^2$ | $\psi_3^4 q_3$ | $\psi_3^3 \cdot \psi_4^2$ | $\psi_3^4 \cdot \psi_5^2$ | $\psi_2^5 \cdot \psi_6^2$ |
| $x_4$ | 1 | $\psi_4 \cdot \psi_2^3$ | $\psi_4^2 \cdot \psi_3^3$ | $\psi_4^6 q_4$ | $\psi_4^4 \cdot \psi_5^3$ | $\psi_4^5 \cdot \psi_6^3$ |
| $x_5$ | 1 | $\psi_5 \cdot \psi_2^4$ | $\psi_5^2 \cdot \psi_3^4$ | $\psi_5^3 \cdot \psi_4^4$ | $\psi_5^8 q_5$ | $\psi_5^5 \cdot \psi_6^4$ |
| $x_6$ | 1 | $\psi_6 \cdot \psi_2^5$ | $\psi_6^2 \cdot \psi_3^5$ | $\psi_6^3 \cdot \psi_4^5$ | $\psi_6^4 \cdot \psi_5^5$ | $\psi_6^1 0 q_6$ |

Warunki A.10 wskazują, że mamy do czynienia z quasi-symetrycznym modelem, ich liczba wynosi $(r-1)(r-2)/2$. Warunek A.11 jest typowy dla modelu wierszowo–

Tabela A.4: Lokalne stosunki szans wyznaczone dla tabeli A.3[a]

| $X\backslash Y$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | $\delta_1^2 q_1 q_2$ | $\frac{\delta_1\delta_2}{q_2}$ | $\delta_1\delta_3$ | $\delta_1\delta_4$ | $\delta_1\delta_5$ |
| 2 | $\frac{\delta_2\delta_1}{q_2}$ | $\delta_2^2 q_2 q_3$ | $\frac{\delta_2\delta_3}{q_3}$ | $\delta_2\delta_4$ | $\delta_2\delta_5$ |
| 3 | $\delta_3\delta_1$ | $\frac{\delta_3\delta_2}{q_3}$ | $\delta_3^2 q_3 q_4$ | $\frac{\delta_3\delta_4}{q_4}$ | $\delta_3\delta_5$ |
| 4 | $\delta_4\delta_1$ | $\delta_4\delta_2$ | $\frac{\delta_4\delta_3}{q_4}$ | $\delta_4^2 q_4 q_5$ | $\frac{\delta_4\delta_5}{q_5}$ |
| 4 | $\delta_5\delta_1$ | $\delta_5\delta_2$ | $\delta_5\delta_3$ | $\frac{\delta_5\delta_4}{q_5}$ | $\delta_5^2 q_5 q_6$ |

[a]W stosunku do tabeli A.3 $\delta_1 = \psi_{i+1}/\psi_i$.

kolumnowego, tj. wynika z niego, że:

$$\frac{\Theta_{41}^{XY}}{\Theta_{51}^{XY}} = \frac{\Theta_{42}^{XY}}{\Theta_{52}^{XY}} = \frac{\delta_4}{\delta_5},$$

przy czym należy podkreślić, że warunki te dotyczą tylko stosunków szans, które nie obejmują komórek związanych z przekątną. W przypadku tabeli o wymiarach 6 na 6 jest jeden taki warunek. Daje się co prawda sformułować analogiczny warunek dla lokalnych stosunków szans położonych po drugiej stronie przekątnej, ale są one następstwem quasi–symetrii tego modelu. Ogólnie jest ich:

$$(r-5) + (r-6) + \ldots + 1 = (r-4)(r-5)/2.$$

Kolejny warunek A.12, można przełożyć w odniesieniu do analizowanej tabeli na dwa założenia:

$$\frac{\left(\Theta_{22}^{XY}/\Theta_{32}^{XY}\right)\left(\Theta_{21}^{XY}/\Theta_{31}^{XY}\right)}{\left(\Theta_{23}^{XY}/\Theta_{33}^{XY}\right)\left(\Theta_{24}^{XY}/\Theta_{34}^{XY}\right)} = \frac{q_2 q_3^2 \frac{1}{q_2}}{\frac{1}{q_3^2 q_4} q_4} = q_3^4 = \left(\frac{\Theta_{31}^{XY}/\Theta_{41}^{XY}}{\Theta_{32}^{XY}/\Theta_{42}^{XY}}\right)^4$$

Podobnie zachodzi:

$$\frac{\left(\Theta_{33}^{XY}/\Theta_{43}^{XY}\right)\left(\Theta_{32}^{XY}/\Theta_{42}^{XY}\right)}{\left(\Theta_{34}^{XY}/\Theta_{44}^{XY}\right)\left(\Theta_{35}^{XY}/\Theta_{45}^{XY}\right)} = q_4^4 = \left(\frac{\Theta_{42}^{XY}/\Theta_{52}^{XY}}{\Theta_{43}^{XY}/\Theta_{53}^{XY}}\right)^4$$

Są dwa warunki tego typu a ogólnie jest ich $(r-4)$. Dla analizowanej tabeli jest jeszcze jeden warunek typu A.13.

$$\left(\frac{\Theta_{13}^{XY}/\Theta_{14}^{XY}}{\Theta_{23}^{XY}/\Theta_{24}^{XY}} \frac{\Theta_{24}^{XY}/\Theta_{25}^{XY}}{\Theta_{34}^{XY}/\Theta_{35}^{XY}}\right)^2 = (q_3 q_4)^2 = \frac{\Theta_{33}^{XY}/\Theta_{34}^{XY}}{\Theta_{23}^{XY}/\Theta_{24}^{XY}}.$$

Ogólnie warunków tego typu jest $(r-5)$. Liczba wszystkich warunków A.10-A.13 określających ten model definiuje jego liczbę stopni swobody, wynosi ona:

$$df = (r-1)(r-2)/2 + (r-5)(r-4)/2 + (r-4) + (r-5) = (r-1)(r-2) - r.$$

Warto podkreślić raz jeszcze, że poszczególne warunki mają zastosowanie gdy liczba komórek jest wystarczająca. Na przykład, gdy mamy do czynienia z tabelą o wymiarach 5 na 5 zastosowanie mają wyłączenie warunki A.10 oraz A.12.

### A.4.5 Modele QD i QDS

Jeśli uwzględnienie specyfiki głównej przekątnej dotyczy asymetrycznej wersji modelu zmiennego dystansu (QD) wówczas odpowiednia hipoteza głosi, że:

$$\Theta_{ab}^{XY} = \Theta_k, \text{ gdzie } k = a - b, |k| > 1, \tag{A.14}$$

$$\Theta_{q(q+1)}^{XY} / \Theta_{(q+1)q}^{XY} = const \tag{A.15}$$

$$\Theta_{mm}^{XY} / \Theta_{(m+1)(m+1)}^{XY} = \Theta_{(m+1)(m+2)}^{XY} / \Theta_{(m-1)m}^{XY} \tag{A.16}$$

dla $q$ takiego, że $1 \leq q \leq r-2$, $2 \leq m \leq r-3$. Ilustrację tego modelu zawiera tabela 3.40. Warto na przykładzie tej tablicy przełożyć warunki A.14–A.16 na relacje pomiędzy konkretnymi stosunkami szans. Założenie A.14 jest analogiczne do modelu zmiennego dystansu w formie oryginalnej, należy jednak zauważyć, że w przypadku tablicy o wymiarach 5 na 5 przekłada się jedynie na dwa ograniczenia, mianowicie:

$$\Theta_{13}^{XY} = \Theta_{24}^{XY} = \frac{s_2^2}{s_1 s_3} \text{ oraz } \Theta_{31}^{XY} = \Theta_{42}^{XY} = \frac{s_{(-2)}^2}{s_{(-1)} s_{(-3)}}.$$

Warunek A.15 dotyczy lokalnych stosunków szans komórek na pseudo–przekątnych położonych najbliżej głównej przekątnej. Nie są one sobie równe, natomiast, jeśli porówna się wielkości stosunków szans poniżej głównej przekątnej do wielkości symetrycznie położonych stosunków szans powyżej głównej przekątnej, to uzyskana proporcja jest stała, tj:

$$\frac{\Theta_{12}^{XY}}{\Theta_{21}^{XY}} = \frac{\Theta_{23}^{XY}}{\Theta_{32}^{XY}} = \frac{\Theta_{34}^{XY}}{\Theta_{43}^{XY}} = \frac{s_1^2 / s_2}{s_{(-1)}^2 / s_{(-2)}}.$$

Ostatni z warunków dotyczy stosunków szans na głównej przekątnej. W odniesieniu do tabeli o wymiarach 5 na 5 przekłada się to na równość:

$$\frac{\Theta_{22}^{XY}}{\Theta_{33}^{XY}} = \frac{\Theta_{34}^{XY}}{\Theta_{12}^{XY}} = \frac{q_2}{q_4}.$$

Bardziej ogólnie, z A.14 wynikają $(r-3)(r-4)$ warunki, założenie A.15 przekłada się na $r-3$ warunki, natomiast z formuły A.16 wynikają $r-4$ warunki. Liczba stopni swobody modelu QD, która zdaje sprawę z łącznej liczby tych warunków wynosi:

$$df = r^2 - 5r + 5.$$

Jeśli modyfikacja dotyczy symetrycznej wersji modelu dystansu (QDS), wówczas do warunków A.14, A.16, należy dodać warunek o symetrii lokalnych stosunków szans, który przekłada się na $(r-2)(r-1)/2$. Zauważmy, że z warunku symetrii wynika warunek A.15, dlatego można go pominąć. Ponadto uwzględniając symetrię lokalnych

stosunków szans można zredukować liczbę niezależnych warunków wynikających A.14 o połowę. Liczba stopni swobody dla tego modelu wynosi:

$$df = (r-2)(r-1)/2 + (r-3)(r-4)/2 + (r-4) = (r-1)(r-3).$$

Podobnie jak przypadku asymetrycznej wersji uwzględnienie specyfiki głównej przekątnej wymaga dodatkowo $r-1$ niezależnych parametrów.

### A.4.6 Model QFD

Zgodnie z modelem QFD lokalne stosunki szans, które nie obejmują komórek głównych przekątnych są równe 1, co daje $(r-3)(r-2)$ warunki. Stosunki szans na pseudo–przekątnych przylegających do głównej przekątnej są symetryczne, tj. $\Theta^{XY}_{q(q+1)} = \Theta^{XY}_{(q+1)q}$ co przekłada się $(r-2)$ warunki. Uwzględnić należy również sformułowane $r-4$ warunki A.16, dotyczące relacji stosunków szans na głównej przekątnej, tak jak miało to miejsce w przypadku modelu zmiennego dystansu D. Przykładowo w odniesieniu do tabeli 3.45 zamieszczonej w rozdziale 3 przekłada się to na równość:

$$\Theta^{XY}_{22}/\Theta^{XY}_{33} = \Theta^{XY}_{34}/\Theta^{XY}_{12}.$$

Liczba stopni swobody tego modelu wynosi więc:

$$df = (r-3)(r-2) + (r-2) + (r-4) = r^2 - 3r.$$

## A.5 Równoważność modeli dla tablicy o wymiarach 4 x 4

### A.5.1 Modele QS i QDS

Jeśli model QDS formułujemy dla tabeli o wymiarach 4 x 4 wówczas warunki A.14 oraz A.16 nie mają zastosowania. Pozostaje warunek głoszący symetrię lokalnych stosunków szans, tj.

$$\Theta^{XY}_{13} = \Theta^{XY}_{31} \text{ oraz } \Theta^{XY}_{12} = \Theta^{XY}_{21} \text{ oraz } \Theta^{XY}_{23} = \Theta^{XY}_{32}.$$

### A.5.2 Modele QCP i QFD

Przypomnijmy, że zgodnie z modelem CP (3.40), lokalne stosunki szans $\Theta^{XY}_{ij}$ są równe 1, o ile $i \neq j$. Uwzględniając specyfikę głównej przekątnej (porównaj tabelę 3.43), warunek ten przestaje być adekwatny dla lokalnych stosunków szans wyznaczonych dla komórek na pseudo–przekątnych do niej przylegających. Stosunki te pozostają

jednak symetryczne. W związku z tym, model QCP dla tabeli o wymiarach 4 x 4, przekłada się na cztery warunki:

$$\Theta_{13}^{XY} = 1,\ \Theta_{31}^{XY} = 1,\ \Theta_{12}^{XY} = \Theta_{21}^{XY},\ \text{oraz}\ \Theta_{23}^{XY} = \Theta_{32}^{XY}.$$

Podobnie jest z modelem QFD. Przypomnijmy, że model FD głosił warunki modelu CP i modelu D, tj. warunek 3.44:

$$\Theta_{ij}^{XY} = 1 \ \text{ oraz } \ \Theta_{qq}^{XY} = const,$$

dla $i \neq j$, i każdego $q$, takiego że $1 \leq q \leq r$. Po uwzględnieniu specyfiki głównej przekątnej (porównaj tabelę 3.45) drugi z wymienionych warunków przestaje być adekwatny, natomiast pierwszy nie dotyczy stosunków szans wyznaczonych dla komórek na pseudo–przekątnych przylegających do głównej przekątnej. Lokalne stosunki szans wyznaczone dla tych komórek są symetryczne — podobnie jak w modelu QCP. Na ogół model QFD, głosi jeszcze warunki A.16, ale mają one zastosowanie dla tabeli o wymiarach co najmniej 5 x 5. Model QFD dla tabeli o wymiarach 4 x 4 będzie głosił te same — przytoczone powyżej — warunki co model QCP.

# Bibliografia


Agresti, A. (1984), *Analysis of Ordinal Categorical Data*, John Wiley, New York.

Agresti, A. (2002), *Categorical Data Analysis*, 2nd edn, John Wiley, New York.

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, John Wiley, New Jersey.

Agresti, A., Chuang, C. i Kezouh, A. (1987), 'Order-restricted score parameters in association models for contingency tables', *Journal of the American Statistical Association* **82**, 619–623.

Alba, R. (1987), 'Interpreting the parameters of log-linear models', *Sociological Methods and Research* **16**, 45–77.

Aldrich, J. i Nelson, F. (1984), *Linear Probability, Logit, and Probit Models*, Sage, Beverly Hills, CA.

Andersen, E. (1980), *Discrete Statistical Models with Social Science Applications*, North-Holland, Amsterdam.

Becker, M. (1990), 'Quasisymmetric models for the analysis of square contingency tables', *Journal of the Royal Statistical Society. Series B (Methodological)* **52**, 369–378.

Birch, M. W. (1963), 'Maximum likelihood in three-way contingency tables', *Journal of the Royal Statistical Society. Series B (Methodological)* **25**, 220–233.

Bishop, Y. M., Fienberg, S. E. i Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge.

Blau, P. i Duncan, O. D. (1967), *The American Occupational Structure*, Academic Press, New York.

Blumen, I., Kogan, M. i McCarthy, P. J. (1955), *The Industrial Mobility of Labor as a Probability Process*, Cornell University Press, Ithaca New York.

Bojanowski, M. (2003), *Strukturalne modele informacyjne i modele logarytmiczno-liniowe. Dwa podejścia do analiz rozkładów zmiennych nominalnych (praca magisterska)*, Uniwersytet Warszawski.

Bonett, D. G. i Bentler, P. M. (1983), 'Goodness–of–fit procedures for the evaluation and selection of log-linear models', *Psychological Bulletin* **93**, 149–166.

Borooah, V. (2001), *Logit and Probit: Ordered and Multinomial Models*, Sage, Thousand Oaks, London.

Boudon, R. (1973), *Mathematical Structures of Social Mobility*, Elsevier, Amsterdam.

Breen, R. (1985), 'Models for the comparative analysis of vertical mobility', *Quality and Quantity* **19**, 337–352.

Breen, R. (red.) (2006), *Social Mobility in Europe*, Oxford University Press, Oxford.

Caussinus, H. (1965), 'Contribution a l'analyse statistique des tableaux de correlation', *Annalles de la Faculte des Sciences de l'Universite de Tolouse* **29**, 77–182.

Cichomski, B. (kierownik projektu), Jerzyński, T. i Zieliński, M. (2009), *Polskie Generalne Sondaże Społeczne, skumulowany komputerowy zbiór danych 1992–2008*, Instytut Studiów Społecznych, Uniwersytet Warszawski.

Clogg, C. (1981), 'Latent structure models of mobility', *American Journal of Sociology* **86**, 836–868.

Clogg, C. (1982*a*), 'Some models for the analysis of association in multiway cross-classifications having ordered categories', *Journal of the American Statistical Associations* **77**, 803–815.

Clogg, C. (1982*b*), 'Using association models in sociological research: Some examples', *American Journal of Sociology* **88**, 114–134.

Clogg, C. (1995), Latent class models, *in* G. Arminger, C. Clogg i S. M. E., (red.) 'Handbook of statistical modeling for the social and behavioral sciences', Plenum Publishing Corporation, New York, s. 311–359.

Cochran, W. G. (1954), 'Some methods of strengthening the common $\chi^2$ tests', *Biometrics* **10**, 417–451.

Czapliński, J. i Panek, T. (red.) (2007), *Diagnoza społeczna. Warunki i jakość życia Polaków 2007*, Vizja Press, Warszawa.

Demaris, A. (1998), *Logit Modelling Practical Applications*, Sage, Bevely Hills, London.

Dobson, A. (2002), *An Introduction to Generalized Linear Models*, 2 edn, Chapman and Haal/CRC, Boca Raton.

Domański, H. (1989), Przemiany struktury społecznej ludności w Polsce, *W:* H. Domański i J. Witkowski, (red.) 'Struktura społeczno–zawodowa a ruchliwość społeczna i przestrzenna ludności w Polsce', Szkoła Główna Planowania i Statystyki, Warszawa, s. 11–110.

Domański, H. (1996), 'Analiza zależności między zmiennymi kategorialnymi. Przykłady zastosowania programu GLIM', *ASK. Społeczeństwo, Badania, Metody* **4**, 103–130.

Domański, H. (2004), *O ruchliwości społecznej w Polsce*, Wydawnictwo IFiS PAN, Warszawa.

Domański, H. (2007*a*), 'New dimension of social stratification in Poland? Class Membership and Electoral Voting in 1991-2001', *European Sociological Review* **2(24)**, 169–182.

Domański, H. (2007*b*), *Struktura społeczna*, Wydawnictwo Naukowe Scholar, Warszawa.

Domański, H., Mach, B. W. i Przybysz, D. (2008), Pochodzenie społeczne — wykształcenie — zawód: ruchliwość społeczna w Polsce w latach 1982–2006, *W:* H. Domański, (red.) 'Zmiany stratyfikacji społecznej w Polsce', Wydawnictwo IFiS PAN, Warszawa, s. 97–132.

Domański, H. i Przybysz, D. (2007), *Homogamia małżeńska a hierarchie społeczne*, Wydawnictwo IFiS PAN, Warszawa.

Domański, H. i Sawiński, Z. (1987), 'Dimensions of occupational mobility: The empirical invariance', *European Sociological Review* **3**, 39–53.

Domański, H. i Sawiński, Z. (1995), 'Polska Społeczna Klasyfikacja Zawodów. PSKZ–1995. Propozycja badawcza', *ASK. Społeczeństwo, Badania, Metody* **4**, 77–94.

Domański, H., Słomczyński, K. M. i Sawiński, Z. (2008), *Nowa Klasyfikacja i Skale Zawodów. Socjologiczne wskaźniki pozycji społecznej w Polsce*, Wydawnictwo IFiS PAN, Warszawa.

Duncan, O. D. (1966), Methodological issues in the analysis of social mobility, *W:* N. J. Smelser i L. S. M., (red.) 'Social Structure and Mobility in Economic Development', Aldine Publishing Company, Chicago, s. 51–97.

Duncan, O. D. (1975), *Introduction to Structural Equation Models*, Academic Press, New York.

Duncan, O. D. (1979), 'How destination depends on origin in the occupational mobility table', *American Journal of Sociology* **88**, 793–804.

Duncan, O. D. (1980), 'Testing key hypotheses in panel analysis', *Sociological Methodology* s. 279–289.

Duncan, O. D. (1981), 'Two faces of panel analysis: parallels with comparative cross-sectional analysis and time-lagged association', *Sociological Methodology* s. 281–318.

Duncan, O. D. (1985), 'New light on the 16-fold table', *American of Journal Sociology* s. 88–128.

Erikson, R. i Goldthorpe, J. H. (1992), *The Constant Flux: A Study of Class Mobility in Industrial Societies*, Oxford University Press, Oxford.

Featherman, D. L. i Hauser, R. M. (1978), *Opportunity and change*, Academic Press, New York.

Featherman, D. L., Jones, F. L. i Hauser, R. M. (1975), 'Assumptions of social mobility research in the US: the case of occupational status', *Social Science Research* **4**, 329–340.

Fienberg, S. E. (1980), *The Analysis of Cross-Classified Categorical Data*, MIT Press, Cambridge.

Fingleton, B. (1984), *Models of Category Counts*, Cambridge University Press, Cambridge.

Fossett, M. A., Galle, O. R. i Kelly, W. R. (1986), 'Racial occupational inequality, 1940-1980: National and regional trends', *American Sociological Review* **51**, 421–429.

Gelman, A. i Rubin, D. B. (1995), 'Avoiding model selection in bayesian social research', *Sociological Methodology* **25**, 165–173.

Gelman, A. i Rubin, D. B. (1999), 'Evaluating and Using Statistical Methods in the Social Sciences. A Discussion of "A Critique of the Bayesian Information Criterion for Model Selection"', *Sociological Methods and Research* **27**, 403–410.

Gini, C. (1914), 'Di una misura della dissomiliglianza tra due gruppi di quan- titŕa e delle sue applicazioni allo studio delle relazioni statistiche', *Reale Instituto Veneto di Scienze, Lettere ed Art (Series 8)* **74**, 185–213.

Glass, D. (red.) (1954), *Social Mobility in Britain*, Routledge and Kegan Paul, London.

Główny Urząd Statystyczny (2004), *Rocznik Demograficzny*, Wydawnictwo GUS, Warszawa.

Goodman, L. A. (1961), 'Statistical methods for the mover-stayer model', *Journal of the American Statistical Association* **56**, 841–868.

Goodman, L. A. (1963), 'Simple methods for analyzing three-factor interaction in contingency tables', *Journal of the American Statistical Associations* **59**, 319–352.

Goodman, L. A. (1965), 'On the statistical analysis of mobility tables', *American Journal of Sociology* **70**, 564–585.

Goodman, L. A. (1969), 'On the measurement of social mobility: An index of status persistence', *American Sociological Review* **34**, 831–850.

Goodman, L. A. (1970), 'The multivariate analysis of qualitative data: Interactions among multiple classifications', *Journal of the American Statistical Associations* **65**, 226–256.

Goodman, L. A. (1971), 'The analysis of multidimensional contingency tables: Stepwise procedures and direct methods for building models for multiple classification', *Technometrics* **13**, 33–61.

Goodman, L. A. (1972*a*), 'A general model for the analysis of surveys', *American Journal of Sociology* **77**, 1035–1086.

Goodman, L. A. (1972*b*), 'A modified multiple regression approach to the analysis of dichotomous variables', *American Sociological Review* **37**, 28–46.

Goodman, L. A. (1972*c*), Some multiplicative models for the analysis of cross-classified data, *W:* L. Cam, J. Neyman i E. Scott, (red.) 'Proceedings of the

Sixth Berkeley Symposium on Mathematical Statistics and Probability', University of California Press, s. 649–696.

Goodman, L. A. (1973), 'The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach', *Biometrika* **60**(1), 179–192.

Goodman, L. A. (1974), 'Exploratory latent structure analysis using both identifiable and unidentifiable models', *Biometrika* **61**(2), 215–231.

Goodman, L. A. (1975), 'The relationship between modified and usual multiple regression approaches to the analysis of dichotomous variables', *Sociological Methodology* **25**, 83–110.

Goodman, L. A. (1979*a*), 'Multiplicative models for the analysis of occupational mobility tables and other kinds of cross-classification tables', *American Journal of Sociology* **84**, 804–819.

Goodman, L. A. (1979*b*), 'Simple models for the analysis of association in cross-classifications having ordered categories', *Journal of the American Statistical Associations* **74**, 537–552.

Goodman, L. A. (1981*a*), 'Criteria for determining whether certain categories in a cross-classification table should be combined, with special reference to occupational categories in an occupational mobility table', *American Journal of Sociology* **87**, 612–650.

Goodman, L. A. (1981*b*), 'Three elementary views of log linear models for the analysis of cross-classifications having ordered categories', *Sociological Methodology* **12**, 193–239.

Goodman, L. A. (1985), 'The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries', *The Annals of Statistics* **13**, 10–69.

Goodman, L. A. (1986), 'Some useful extensions of the usual correspondence analysis approach and the usual log-linear approach in the analysis of contingency tables', *International Statistical Review* **54**, 254–309.

Goodman, L. A. (2007), 'Statistical magic and/or statistical serendipity: An age of progress in the analysis of categorical data', *Annual Review of Sociology* **33**, 1–19.

Goodman, L. A. i Hout, M. (1998), 'Statistical methods and graphical displays for analyzing how the association between two qualitative variables differs among countries, among groups, or over time: A modified regression-type approach', *Sociological Methodology* **28**, 175–230.

Goodman, L. A. i Hout, M. (1998b), 'Rejoinder: Understanding the Goodman-Hout approach to the analysis of differences in association and some related comments', *Sociological Methodology* **28**, 249–261.

Goodman, L. A. i Hout, M. (2001), 'Statistical Methods and Graphical Displays for Analyzing How the Association between Two Qualitative Variables Differs among Countries, among Groups, or over Time. Part II: Some Exploratory Techniques, Simple Models, and Simple Examples', *Sociological Methodology* **31**, 189–221.

Goodman, L. A. i Kruskal, W. H. (1959), 'Measures of Association for Cross Classifications II: Further Discussion and References', *Journal of the American Statistical Association* **54**, 123–163.

Haberman, S. J. (1973), 'The analysis of residuals in cross-classified tables', *Biometrics* **29**, 205–220.

Haberman, S. J. (1974*a*), 'Log-linear models for frequency tables with ordered classifications', *Biometrics* **30**, 589–600.

Haberman, S. J. (1974*b*), *The Analysis of Frequency Data*, University of Chicago Press Chicago.

Haberman, S. J. (1978), *Analysis of Qualitative Data. Vol. 1. Introductory Topics*, Academic Press, New York.

Haberman, S. J. (1979), *Analysis of Qualitative Data. Vol. 2. New Developments*, Academic Press, New York.

Haberman, S. J. (1981), 'Tests for independence in two-way contingency tables based on canonical correlation and on linear–by–linear interaction', *The Annals of Statistics* **9**, 1178–1186.

Hagenaars, J. (1990), *Categorical Longitudinal Data. Log-Linear Panel, Trend and Cohort Analysis*, Sage, Newbury Park London New Delhi.

Hagenaars, J. (1993), *Loglinear Models with Latent Variables*, Sage, Thousand Oaks, London.

Hagenaars, J. i McCutcheon, A. (2002), *Applied Latent Class Analysis*, Cambridge University Press Cambridge.

Halpin, B. i Chan, T. (2003), 'Educational homogamy in Ireland and Britain: trends and patterns', *British Journal of Sociology* **54**(4), 473–496.

Hauser, R. M. (1980), 'Some exploratory methods for modelling mobility tables and other cross-classified data', *Sociological Methodology* **11**, 413–458.

Hildebrand, D. K., Laing, J. D. i Rosenthal, H. (1977), *Analysis of Ordinal Variables*, Sage, Bevely Hills, London.

Hope, K. (1981), 'The new mobility ratio', *Social Forces* **60**, 544.

Hout, M. (1983), *Mobility Tables*, Sage, Beverly Hills.

Ishii-Kuntz, M. (1994), *Ordinal Log-Linear Models*, Sage, Thousand Oaks.

Janicka, K. (1973), Ruchliwość międzypokoleniowa, *W:* K. M. Słomczyński, W. Wesołowski i B. Mach, (red.) 'Struktura i ruchliwość społeczna', Ossolineum, Wrocław, s. 61–101.

Janicka, K. (1976), *Ruchliwość społeczno-zawodowa i jej korelaty*, Ossolineum, Wrocław.

Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press, Oxford.

Kahl, J. A. (1957), *The American Class Structure*, Irvington Publishers, New York.

Kalmijn, M. (1991), 'Status homogamy in the United States', *American Journal of Sociology* **97**, 496–523.

Karpiński, J. (2006), *Wprowadzenie do metodologii nauk społecznych*, WSPZ im. Leona Koźmińskiego, Warszawa.

Kendall, M. G. (1948), *Rank Correlation Methods*, Griffin, London.

Klatzky, S. R. i Hodge, R. W. (1971), 'A canonical correlation analysis of occupational mobility', *Journal of the American Statistical Association* **66**, 16–22.

Knoke, D. i Burke, P. (1980), *Log-Linear Models*, Sage, Bevely Hills, London.

Kruskal, W. H. (1958), 'Ordinal measures of association', *Journal of the American Statistical Association* **53**, 814–861.

Kuha, J. i Firth, D. (2004), *On the Index of Dissimilarity for Lack of Fit in Log–linear and Log–multiplicative Models*, http://stats. lse. ac. uk/kuha/Publications/publications.html.

Kutylowski, J. (1979), *On Alternative Parametrisations in Standard Log-linear Model*, Institute für Höhere Studien, Wien.

Kutyłowski, A. J. (1988), 'Structural analysis of social mobility tables', *Paper presented at the Workshop "Labor Mobility in Nordic Countries" organized by the Trade Union Institute for Economic Research, Stockholm* .

Kutyłowski, A. J. (1989), Analysis of symmetric cross-classifications, *W:* A. Decarli, B. Francis, R. Gilchrist i G. Seeber, (red.) 'Statistical Modelling', Springer, New York, s. 188–197.

Kutyłowski, Andrzej, J. (1994), Analysis of ordinal uniresponse data with log-linear models, *W:* S. Boelskifte, (red.) 'Symposium i Anvendt. Statistik', Kopenhaga, s. 344–358.

Lawal, H. B. (2003), 'The structure of the log odds-ratios in non-independence and symmetry diagonal models for square contingency tables', *Quality and Quantity* **37**(2), 111–134.

Lawal, H. B. (2004), 'Review of non-independence, asymmetry, skew-symmetry and point-symmetry models in the analysis of social mobility data', *Quality and Quantity* **38**(3), 259–289.

Lazarsfeld, P. F. (1950), The logical and mathematical structure of latent structure analysis, *W:* S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld, S. Star i J. Clausen, (red.) 'Measurement and prediction, t. 4', Princeton University Press Princeton, New Jersey, s. 362–412.

Lazarsfeld, P. F. i Henry, N. W. (1968), *Latent Structure Analysis*, Houghton, Mifflin.

Liao, T. (1994), *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*, Sage, Thousand Oaks, London.

Lieberson, S. (1976), 'Rank-sum comparisons between groups', *Sociological Methodology* **7**, 276–291.

Lissowski, G. (1978), *Zależności statystyczne między uporządkowaniami*, WSNS, Warszawa.

Lissowski, G. (1984), 'Zastosowanie modeli logarytmiczno-liniowych do analizy związków między wieloma zmiennymi jakościowymi', *Studia Socjologiczne* **2**, 239–263.

Lissowski, G. (1991), 'Dekompozycja tablic ruchliwości społecznej', *Przegląd Socjologiczny* **39**, 167–210.

Lissowski, G., Haman, J. i Jasiński, M. (2008), *Podstawy statystyki dla socjologów*, Wydawnictwo Naukowe Scholar, Warszawa.

Liu, I. i Agresti, A. (2005), 'The analysis of ordered categorical data: an overview and a survey of recent developments', *Test* **14**(1), 1–73.

Mach, B. W. (2002), Patterns of intergenerational mobility. The long term trends, *W:* K. M. Słomczyński, (red.) 'Social Struture: Changes and Linkages', Wydawnictwo IFiS PAN, Warszawa, s. 29–43.

Mach, B. W. (2004), Intergenerational mobility in Poland: 1972-1988-1994, *W:* R. Breen, (red.) 'Social Mobility in Europe', Oxford University Press, Oxford, s. 269–286.

Mare, R. (1991), 'Five decades of educational assortative mating', *American Sociological Review* **56**, 15–32.

Markus, G. (1984), *Analyzing Panel Data*, Sage, Bevely Hills, London.

Matras, J. (1961), 'Differential fertility, intergenerational occupational mobility, and change in the occupational distribution: some elementary interrelationships', *Population Studies* **15**, 187–197.

McCutcheon, A. L. (1987), *Latent Class Analysis*, Sage, Thousand Oaks, London.

Miller, S. M. (1960), 'Comparative social mobility current sociology', *Current Sociology* **9(1)**, 1–89.

Misztal, M. (1982), *Zróżnicowanie systemu wartości społeczeństwa polskiego*, IS UW, Warszawa.

Misztal, M. (1990), *Elementy systemu wartości współczesnego społeczeństwa polskiego*, PWN, Warszawa.

Nawojczyk, M. i McCutcheon, A. L. (1996), 'Rozdźwięk z z doktryną. postawy amerykańskich i polskich katolików świeckich wobec legalności aborcji', *Studia Socjologiczne* **1**, 49–76.

O'Connell, A. (2006), *Logistic Regression Models for Ordinal Response Variables*, Sage, Thousand Oaks, London.

Pampel, F. (2006), *Logistic Regression: A Primer*, Sage, Thousand Oaks, London.

Pohoski, M. (1983), 'Ruchliwość społeczna a nierówności społeczne', *Kultura i Społeczeństwo* **27**, 135–164.

Pohoski, M. (1991), Społeczne aspekty doboru małżeńskiego w Polsce, *W:* Z. Tyszka, (red.) 'Rodziny polskie o różnym statusie społecznym i środowiskowym', Poznań: Centralny Program Badań Podstawowych, s. 19–42.

Pohoski, M. i Mach, B. W. (1986), Rozmiary i kierunki ruchliwości społecznej w latach 1972-1982, *W:* I. Białecki, (red.) 'Przemiany ruchliwości społecznej w Polsce', Wydawnictwo IFiS PAN, Warszawa, s. 15–49.

Pohoski, M. i Słomczyński, K. M. (1978), *Społeczna Klasyfikacja Zawodów*, Wydawnictwo IFiS PAN, Warszawa.

Pohoski, M., Słomczyński, K. M. i Milczarek, M. (1974), *Społeczna Klasyfikacja Zawodów*, Wydawnictwo IFiS PAN, Warszawa.

Przybysz, D. (2003), 'Modele logarytmiczno–liniowe dla zmiennych porządkowych. przykłady zastosowania', *ASK. Społeczeństwo, Badania, Metody* **13**, 43–47.

Pullum, T. (1975), *Measuring Occupational Inheritance*, Elsevier Press, Amsterdam.

Raftery, A. (1986*a*), 'Choosing models for cross-classifications', *American Sociological Review* **51**, 145–146.

Raftery, A. (1986*b*), 'A note on bayes factors for log-linear contingency table models with vague prior information', *Journal of the Royal Statistical Society. Series B* **48**, 249–250.

Raftery, A. (1995), 'Bayesian model selection in social research', *Sociological Methodology* **25**, 111–163.

Raftery, A. (1999), 'Bayes Factors and BIC. Comment on "A Critique of the Bayesian Information Criterion for Model Selection"', *Sociological Methods and Research* **27**, 411–427.

Raftery, A. i Kass, R. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**, 773–795.

Ritov, Y. i Gilula, Z. (1991), 'The order-restricted rc model for ordered contingency tables: Estimation and testing for fit', *The Annals of Statistics* **19**(4), 2090–2101.

Rogoff, N. (red.) (1953), *Recent Trends in Occupational Mobility*, Free Press, Glencoe.

Sarapata, A. (1965), *Studia nad uwarstwieniem i ruchliwością społeczna w Polsce*, Książka i Wiedza, Warszawa.

Sawiński, Z. (1981), 'Mierniki ruchliwości społeczno–zawodowej', *Studia Socjologiczne* **2**, 171–88.

Smits, J., Ultee, W. i Lammers, J. (1998), 'Educational homogamy in 65 countries: An explanation of differences using country explanatory variables', *American Sociological Review* **63**, 264–285.

Sobel, M. E., Hout, M. i Duncan, O. D. (1985), 'Exchange, structure, and symmetry in occupational mobility', *American Journal of Sociology* **81**, 359–372.

Słomczyński, K. M. (1973), Rola wykształcenia w procesie ruchliwości wewnątrzpokoleniowej, *W:* K. M. Słomczyński i W. Wesołowski, (red.) 'Struktura i ruchliwość społeczna', Ossolineum, Wrocław, s. 61–101.

Słomczyński, K. M., Białecki, I., Domański, H., Janicka, K., Mach, B. W., Sawiński, Z., Sikorska, J. i Wojciech, Z. (1989), *Struktura Społeczna: schemat teoretyczny i warsztat badawczy*, Wydawnictwo IFiS PAN, Warszawa.

Słomczyński, K. M. i Krauze, T. K. (1986), 'Matrix representation of structural and circulation mobility', *Sociological Methods and Research* **14(3)**, 247–269.

Sorokin, P. (1927), *Social and Cultural Mobility*, Free Press Glencoe, Ill.

Styczeń, M. (1989), Dynamika opinii, *W:* S. Nowak, (red.) 'Ciągłość i zmiana tradycji kulturowej', PWN, Warszawa, s. 374–418.

Sztabiński, P. B. i Sztabiński, F. (2003), *Europejski Sondaż Społeczny*, Ośrodek Realizacji Badań Socjologicznych, Instytut Filozofii i Socjologii Polskiej Akademii Nauk, Warszawa. Projekt badawczy finansowany ze środków Komitetu Badań Naukowych.

Tomaszewski, W. (2004), *Modelowanie log-liniowe. Przykłady zastosowań w analizie danych panelowych (praca magisterska)*, Uniwersytet Warszawski.

Tumin, M. M. i Feldman, A. S. (1957), 'Theory and measurement of occupational mobility', *American Sociological Review* **22**, 281–288.

Vermunt, J. K. (1997), *LEM: A general program for the analysis of categorical data*, http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html, Tilburg University.

Wagner, O. (2003), *Analiza klas ukrytych Paula F. Lazarsfelda. Wybrane zagadnienia (praca magisterska)*, Uniwersytet Warszawski.

Weakliem, D. (1999*a*), 'A critique of the bayesian information criterion for model selection', *Sociological Methods and Research* **27**, 359–397.

Weakliem, D. (1999*b*), 'Reply to Firth and Kuha, Gelman and Rubin, Raftery, and Xie', *Sociological Methods and Research* **27**, 436–443.

Weakliem, D. (2004), 'Introduction to the special issue on model selection', *Sociological Methods and Research* **33**(2), 167–187.

Wilensky, H. L. (1966), *Measures and Effects of Social Mobility*, Univ. of California, Inst. of Industrial Relations, Chicago.

Wilks, S. S. (1935), 'The likelihood of independence in contingency tables', *Annals of Mathematical Statistics* **6**, 190–196.

Wilks, S. S. (1938), 'The large-sample distribution of the likelihood for testing composite hypotheses', *Annals of Mathematical Statistics* **9**(1), 60–62.

Xie, Y. (1992), 'The log–multiplicative layer effect model for comparing mobility tables', *American Sociological Review* **57**(3), 380–395.

Xie, Y. (1999), 'The tension between generality and accuracy', *Sociological Methods and Research* **27**, 428–435.

Yamaguchi, K. (1987), 'Models for comparing mobility tables: Toward parsimony and substance', *American Sociological Review* **52**(4), 482–494.

Yamaguchi, K. (1990), 'Some models for the analysis of asymmetric association in square contingency tables with ordered categories', *Sociological Methodology* **20**, 181–212.

Yasuda, S. (1964), 'A methodological inquiry into social mobility', *American Sociological Review* **29**, 16–23.

Zahn, D. A. i Fein, S. B. (1979), 'Large contingency tables with large cell frequencies: A model search algorithm and alternative measures of fit', *Psychological Bulletin* **86**, 1189–1200.