Załącznik 3 Opis modeli log-liniowych wykorzystanych w pracy badawczej„Cudzoziemcy na krajowym rynku pracy w ujęciu regionalnym''

1. Wstęp

W załączniku nr 2 raportu pt. „Cudzoziemcy na krajowym rynku pracy w ujęciu regionalnym'' dokonano kompleksowego przeglądu literatury z zakresu szacowania populacji trudnych do zbadania. Przykład takich populacji stanowić mogą osoby bezdomne, osoby zażywające narkotyki czy cudzoziemcy. Główna trudność związana z tego typu populacjami polega na tym, że brakuje oficjalnych rejestrów bądź innych źródeł statystycznych, które umożliwiałyby estymację liczby osób w tego typu populacjach z akceptowalną precyzją. W literaturze przedmiotu istnieją jednak odpowiednie metody statystyczne, które umożliwiają estymację wielkości populacji trudnych do zbadania. Zaliczyć tutaj można przykładowo rozwiązania bazujące na technikach wielokrotnych połowów (capture-recapture) wykorzystujących modele log-liniowe, analizie klas ukrytych czy modelowaniu z uwzględnieniem tzw. modeli mieszanych z efektami losowymi. Skuteczne wykorzystanie tych technik w praktyce w dużej mierze zależy od dostępności danych statystycznych i jest uwarunkowane koniecznością spełnienia odpowiednich założeń leżących u podstaw poszczególnych metod.

W prezentowanym końcowym raporcie badawczym ostatecznie zdecydowano się skorzystać z modeli log-liniowych na potrzeby estymacji cudzoziemców w Polsce na rynku pracy w ujęciu regionalnym. Wynikało to przede wszystkim z dostępności odpowiednich źródeł danych, które można wykorzystać w tego typu estymacji, dostępności odpowiednich pakietów programu R, w których zaimplementowane są właściwe funkcje na potrzeby modeli log-liniowych oraz kodów na procedurę bootstrap umożliwiającą znalezienie właściwych przedziałów ufności a także z faktu, że w literaturze przedmiotu właśnie te modele są z powodzeniem wykorzystywane w estymacji liczebności populacji trudnych do zbadania. Przykład stanowić mogą prace Coumansa i inn. (2017) oraz Petera i inn. (2012). W pierwszej ze wspomnianych prac wykorzystując trzy źródła danych i modele log-liniowe dokonano oszacowania liczby bezdomnych osób w Holandii z uwzględnieniem dodatkowych przekrojów zdefiniowanych przez płeć, wiek czy pochodzenie. W drugim z przytoczonych artykułów zastosowanie modeli log-liniowych oraz tzw. pasywnych i aktywnych zmiennych pomocniczych umożliwiło oszacowanie liczby osób urodzonych na Bliskim Wschodzie a przebywających w Holandii. Obydwie ze wspomnianych populacji, dla których dokonano szacunków wraz z podaniem odpowiednich 95% przedziałów ufności należą do tzw. populacji trudnych do zbadania.

2. Modele log-liniowe w ujęciu teoretycznym

Modele log-liniowe stanowią obecnie bardzo ważną metodę analizy danych zawartych w tablicach kontyngencji. Rozwój metodologii dedykowanej tej technice analizy danych jakościowych zapoczątkowany został w latach 60-tych XX wieku. Goodman (1964, 1968, 1969) był jednym z pierwszych badaczy, którzy spopularyzowali modele log-liniowe w naukach społecznych. Modele te są szczególnie przydatne w sytuacjach, gdy brak jest precyzyjnego rozróżnienia między zmienną objaśnianą a zmiennymi objaśniającymi, a zachodzi potrzeba wykrycia zależności w pewnym zbiorze danych. Geneza modeli log-liniowych została szczegółowo omówiona również w polskiej literaturze (Brzezińska, 2015).

Punktem wyjścia w modelach log-liniowych jest tablica kontyngencji. Na potrzeby pracy badawczej rozpatrywana będzie jedynie tablice dwuwymiarowe postaci 2×2 oraz trójwymiarowe tablice typu $2 \times 2 \times 2$, aczkolwiek teorię modeli log-liniowych można rozszerzyć na tablice kontyngencji o dowolnych wymiarach. Przedstawione rozważania teoretyczne w głównej mierze bazować będą na pracach Goodmana (1964, 1968, 1969), Stokesa i inn. (2012) oraz Brzezińskiej (2015).

Rozważony będzie na początku przypadek dwudzielczej tablicy kontyngencji 2×2. Załóżmy, że jesteśmy zainteresowani poszukiwaniem zależności między dwiema zmiennymi X i Y mierzonymi na słabych skalach pomiaru, z których każda ma dwa możliwe warianty. Tabela 1 przedstawia łączny rozkład obydwu cech, przy czym n_{ij} oznacza znane liczebności empiryczne na przecięciu i –tego wiersza oraz j –tej kolumny $(i, j = 1{,}2)$, n oznacza całkowitą liczbę jednostek w tablicy dwudzielczej, a n_{i+} oraz n_{+j} odpowiednie liczebności brzegowe, gdzie $n = \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij}$ oraz $n_{i+} = \sum_{j=1}^{2} n_{ij}$ i $n_{+j} = \sum_{i=1}^{2} n_{ij}$.

TABELA 1. DWUDZIELCZA TABLICA KONTYNGENCJI (2×2)

	Kategorie zmiennej Y		
Kategorie zmiennej X	Y_1	Y_2	Razem
X_1	n_{11}	n_{12}	n_{1+}
X_2	n_{21}	n_{22}	n_{2+}
Razem	n_{+1}	n_{+2}	n

Informacje zawarte w Tabeli 1 można również przedstawić wykorzystując w tym celu prawdopodobieństwo $p_{ij} = \frac{n_{ij}}{n}$, takie że $\sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1,\ (i, j = 1{,}2)$, por. Tabela 2.

TABELA 2. DWUDZIELCZA TABLICA KONTYNGENCJI (2×2) PRAWDOPODOBIEŃSTWA

	Kategorie zmiennej Y		
Kategorie zmiennej X	Y_1	Y_2	Razem
X_1	p_{11}	p_{12}	p_{1+}
X_2	p_{21}	p_{22}	p_{2+}
Razem	p_{+1}	p_{+2}	1

Uzasadnieniem do konstrukcji modeli log-liniowych jest niezależność statystyczna, która może być wyrażona w postaci kombinacji liniowej logarytmów odpowiednich prawdopodobieństw. W szczególnym przypadku dwudzielczych tablic kontyngencji 2×2, jeśli zmienne X i Y są statystycznie niezależne spełniony jest następujący warunek:

$$p_{ij} = p_{i+} p_{+j}, \tag{1}$$

dla $i, j = 1{,}2$.

Ponieważ $p_{+1} = p_{11} + p_{21}$ oraz $p_{+2} = p_{12} + p_{22}$ to uwzględniając poniższe związki (przy założeniu statystycznej niezależności rozpatrywanych zmiennych):

$$\frac{p_{11}}{p_{+1}} = \frac{p_{12}}{p_{+2}} = p_{1+} \tag{2}$$

otrzymujemy, że:

$$\frac{p_{11}}{p_{11} + p_{21}} = \frac{p_{12}}{p_{12} + p_{22}} \tag{3}$$

a stąd $p_{11}(p_{12} + p_{22}) = p_{12}(p_{11} + p_{21})$, co po prostych uproszczeniach prowadzi do następującej równości:

$$p_{11}p_{22} = p_{12}p_{21}. \tag{4}$$

Równość (4) orzeka, że jeśli zmienne X i Y są statystycznie niezależne to spełniony jest warunek:

$$\psi = \frac{p_{11}p_{22}}{p_{12}p_{21}} = 1, \tag{5}$$

gdzie ψ to tzw. iloraz szans. Logarytmując obustronnie równość (5) oraz korzystając z podstawowych działań na logarytmach uzyskujemy następującą równość:

$$ln\psi = lnp_{11} - lnp_{12} - lnp_{21} + lnp_{22} = 0. \tag{6}$$

Model log-liniowy dla tablicy dwudzielczej typu 2×2 bierze pod uwagę logarytm ilorazu szans w szczególny sposób. Mianowicie tzw. nasycony model log-liniowy (*ang. saturated model*) definiuje się, jako:

$$ln(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad i,j = 1,2 \tag{7}$$

gdzie $m_{ij} = np_{ij}$ to tzw. liczebności oczekiwane (teoretyczne) w dwudzielczej tablicy kontyngencji odpowiadającej komórce na przecięciu i –tego wiersza i j –tej kolumny. Model ten w przypadku notacji nawiasowej często spotykanej w odniesieniu do modeli log-liniowych można zapisać jako [XY] – model pełny. Powyższy model jest podobny do modelu dwuczynnikowej analizy wariancji dla zmiennej ciągłej y postaci:

$$E(y_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \tag{8}$$

gdzie μ oznacza średni poziom badanej cechy, α_i, β_j to tzw. efekty główne a $(\alpha\beta)_{ij}$ to efekt interakcji. Występujące w równaniu (7) wyrażenia $\mu, \lambda_i^X, \lambda_i^Y, \lambda_{ij}^{XY}$ odnoszą się do standardowej notacji używanej w modelach log-liniowych, przy czym μ to średnia arytmetyczna zlogarytmowanych liczebności rzeczywistych w dwudzielczej tablicy kontyngencji, λ_i^X mierzy wpływ zmiennej X, λ_i^Y mierzy wpływ zmiennej Y, a λ_{ij}^{XY} mierzy wpływ interakcji zmiennych X i Y. Wartości teoretyczne m_{ij} dla pełnego modelu log-liniowego przedstawia Tabela 3.

TABELA 3. DWUDZIELCZA TABLICA KONTYNGENCJI (2×2) – WARTOŚCI TEORETYCZNE

	Kategorie zmiennej Y	
Kategorie zmiennej X	Y_1	Y_2
X_1	$e^{(\mu+\lambda_1^X+\lambda_1^Y+\lambda_{11}^{XY})}$	$e^{(\mu+\lambda_1^X-\lambda_1^Y-\lambda_{11}^{XY})}$
X_2	$e^{(\mu-\lambda_1^X+\lambda_1^Y-\lambda_{11}^{XY})}$	$e^{(\mu-\lambda_1^X-\lambda_1^Y+\lambda_{11}^{XY})}$

Parametry modelu (7) możemy z kolei wyrazić następującymi wzorami:

$$\mu = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2}\ln(m_{ij})}{4}, \tag{9}$$

$$\lambda_i^X = \frac{\sum_{j=1}^{2}\ln(m_{ij})}{2} - \mu, \tag{10}$$

$$\lambda_j^Y = \frac{\sum_{i=1}^{2} \ln(m_{ij})}{2} - \mu, \tag{11}$$

$$\lambda_{ij}^{XY} = \ln(m_{ij}) - \lambda - \lambda_i^X - \lambda_j^Y. \tag{12}$$

Iloraz szans (5) może być również wyrażony, jako funkcja liczebności oczekiwanych:

$$\psi = \frac{m_{11}m_{22}}{m_{12}m_{22}} \tag{13}$$

tak więc biorąc pod uwagę liczebności teoretyczne zamieszczone w Tabeli 3. po prostych przekształceniach otrzymujemy, że:

$$ln\psi = lnm_{11} - lnm_{12} - lnm_{21} + lnm_{22} = 4\lambda_{11}^{XY}. \tag{14}$$

W takiej sytuacji hipoteza zerowa o niezależności zmiennych X i Y jest równoważna hipotezie $H_0: \lambda_{11}^{XY} = 0$. Na potrzeby weryfikacji tej hipotezy można wykorzystać statystykę testową G^2 postaci:

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \ln\left(\frac{n_{ij}}{\widehat{m}_{ij}}\right), \tag{15}$$

która w przypadku, gdy H_0 jest prawdziwa ma asymptotyczny rozkład χ^2 z 1 stopniem swobody, przy czym $\widehat{m}_{ij} = n_{i+}n_{+j}/n$ stanowią oszacowania liczebności teoretycznych wyznaczonych dla danego modelu log-liniowego.

W powyższym przypadku mamy do czynienia z tzw. modelem niezależności (*ang. independence model*), który wyraża się wzorem:

$$ln(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y, \tag{16}$$

i zawiera jedynie parametry wpływu pojedynczych zmiennych. W notacji nawiasowej model ten możemy zapisać jako [X][Y]. Model ten jest kombinacją liniową trzech parametrów wpływu, w którym występuje efekt główny i efekty badanych zmiennych, natomiast brak jest w nim wpływu interakcji.

Można również rozpatrywać inne modele log-liniowe. Przykładowo w przypadku dwóch zmiennych można zbudować model z jedną zmienną postaci X (w notacji nawiasowej [X]):

$$ln(m_{ij}) = \mu + \lambda_i^X, \tag{17}$$

który uwzględnia jedynie wpływ tej zmiennej. Analogicznie można rozpatrywać model dla zmiennej Y (w notacji nawiasowej [Y]), który bierze pod uwagę wpływ zmiennej Y i może być przedstawiony w postaci:

$$ln(m_{ij}) = \mu + \lambda_i^Y. \tag{18}$$

Istnieje również możliwość zbudowania modelu zerowego ([0]), w którym brak jest wpływu jakiejkolwiek zmiennej:

$$ln(m_{ij}) = \mu. \tag{19}$$

W teorii modeli log-liniowych buduje się zazwyczaj wiele modeli, które są ze sobą porównywane (przykładowo w rozpatrywanym przypadku tablicy 2×2 można rozpatrywać cztery modele). Model pełny zawiera wpływ wszystkich zmiennych i z punktu widzenia złożoności jest najbardziej skomplikowanym modelem. Celem jest zatem poszukanie modelu prostszego w stosunku do modelu pełnego, który jednocześnie byłby dobrze dopasowany do danych.

Rozważmy obecnie trójwymiarową tablicę kontyngencji $2 \times 2 \times 2$, pokazaną w Tabeli 4 dla trzech zmiennych X, Y, Z, z których każda przyjmuje dwa warianty.

TABELA 4. TRÓJDZIELCZA TABLICA KONTYNGENCJI $(2 \times 2 \times 2)$

Kategorie zmiennej X	Kategorie zmiennej Z				Razem
	Z_1		Z_2		
	Kategorie zmiennej Y		Kategorie zmiennej Y		
	Y_1	Y_2	Y_1	Y_2	
X_1	n_{111}	n_{121}	n_{112}	n_{122}	n_{1++}
X_2	n_{211}	n_{221}	n_{212}	n_{222}	n_{2++}
Razem	n_{+11}	n_{+21}	n_{+12}	n_{+22}	n

Modeł pełny – nasycony (*ang. saturated model*) dla trzech zmiennych w postaci addytywnej jest postaci:

$$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + +\lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}\,, \tag{20}$$

przy czym jego paramtery dla tablicy $2 \times 2 \times 2$ wyrażają się poniższymi wzorami:

$$\mu = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\ln(m_{ijk})}{8}, \tag{21}$$

$$\lambda_i^X = \frac{\sum_{j=1}^{2}\sum_{k=1}^{2}\ln(m_{ijk})}{4} - \mu, \tag{22}$$

$$\lambda_j^Y = \frac{\sum_{i=1}^{2}\sum_{k=1}^{2}\ln(m_{ijk})}{4} - \mu, \tag{23}$$

$$\lambda_k^Z = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2}\ln(m_{ijk})}{4} - \mu, \tag{24}$$

$$\lambda_{ij}^{XY} = \frac{\sum_{k=1}^{2}\ln(m_{ijk})}{2} - \lambda_i^X - \lambda_j^Y - \mu, \tag{25}$$

$$\lambda_{ik}^{XZ} = \frac{\sum_{j=1}^{2}\ln(m_{ijk})}{2} - \lambda_i^X - \lambda_k^Z - \mu, \tag{26}$$

$$\lambda_{jk}^{YZ} = \frac{\sum_{i=1}^{2}\ln(m_{ijk})}{2} - \lambda_j^X - \lambda_k^Z - \mu, \tag{27}$$

$$\lambda_{ij}^{XY} = \ln(m_{ijk}) - \lambda_i^X - \lambda_j^Y - \lambda_k^Z - \lambda_{ij}^{XY} - \lambda_{ik}^{XZ} - \lambda_{jk}^{YZ} - \mu, \tag{28}$$

przy czym m_{ijk} oznacza wartości teoretyczne w trójdzielczej tablicy kontyngencji. Podobnie jak w przypadku modeli dla dwuwymiarowych tablic kontyngencji można rozpatrywać i porównywać ze sobą wiele różnych modeli log-liniowych. Odbywa się to według zasady hierarchiczności, według której parametry niższego rzędu nie mogą być usunięte z modelu, dopóki parametr wyższego rzędu będzie zawierał jakąkolwiek zmienną występującą w parametrze niższego rzędu. W tym przypadku mamy do czynienia z tzw. hierarchicznymi modelami log-liniowymi. W sytuacji trzech zmiennych X, Y, Z można zbudować dziewięć różnych modeli opisujących związki pomiędzy rozpatrywanymi zmiennymi. Ich

równania oraz zapis nawiasowy przedstawiono w Tabeli 5.

W analizie log-liniowej głównym celem jest wybór modelu o możliwie najprostszej postaci, który jednocześnie byłby najlepiej dopasowany do danych. W literaturze przedmiotu (Goodman 1964, 1968, 1969; Stokes i inn. 2012; Brzezińska 2015) proponuje się na potrzeby oceny modeli różnego rodzaju kryteria. Zostały one również wykorzystane w niniejszej pracy badawczej w wyborze finalnego modelu. Do najważniejszych kryteriów zaliczamy iloraz wiarogodności, dewiancja, AIC, BIC oraz współczynnik pseudo R^2.

Iloraz wiarogodności jest miarą pozwalającą ocenić dopasowanie modelu do danych. Przykładowo dla tablic 2×2 wyraża się on wzorem (15). W sytuacji gdy wartość ilorazu wiarogodności G^2 jest duża to wówczas model taki powinien być odrzucony jako model, który w nieprawidłowy sposób odwzorowuje zależności między badanymi zmiennymi. Współczynnik G^2 może być także wykorzystywany do porównania oceny różnych modeli. W sytuacji gdy porównujemy dwa modele współczynnik G^2 może zostać przedstawiony w postaci (dla tablic 2×2) :

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} \widehat{m}_{ij}^{0} \ln\left(\frac{\widehat{m}_{ij}^{0}}{\widehat{m}_{ij}^{1}}\right), \tag{29}$$

gdzie: 0 odnosi się do liczebności teoretycznych modelu ogólniejszego, tj. zawierającego wszystkie możliwe parametry, natomiast 1 dotyczy liczebności teoretycznych modelu zagnieżdżonego o uproszczonej postaci i zawierającego się w modelu 0. Współczynnik ten może być również przedstawiony w postaci:

$$G^2(M_0|M_1) = G^2(M_0) - G^2(M_1). \tag{30}$$

Powyższa statystyka ma rozkład chi-kwadrat o liczbie stopni swobody $df = df(M_0) - df(M_1)$ gdzie M_0 jest modelem zagnieżdżonym, a M_1 modelem ogólnym z większą liczbą parametrów i nazywana jest dewiancją. Dewiancja pozwala ocenić czy parametr występujący w modelu M_1, a niewystępujący w modelu M_0 jest statystycznie istotny.

Statystyką służącą do porównywania ze sobą większej liczby modeli jest tzw. kryterium informacyjne Akaike oraz Schwarza (bayesowskie). Kryterium informacyjne Akaike wyraża się wzorem:

$$AIC = G^2 - df, \tag{31}$$

gdzie G^2 to iloraz wiarogodności badanego modelu a df to liczba odpowiadających mu stopni swobody. Z kolei bayesowskie kryterium informacyjne wyraża się wzorem:

$$BIC = G^2 - df \cdot ln(n), \tag{32}$$

gdzie n to liczebność w tablicy kontyngencji. Preferowane są przy tym modele, dla których miary AIC i BIC przyjmują mniejsze wartości.

W przypadku modeli log-liniowych do ich oceny wykorzystuje się również tzw. współczynnik determinacji R^2 (pseudo R^2). Wyraża się on wzorem:

$$R^2 = \frac{G^2(M_0) - G^2(M_1)}{G^2(M_0)}. \tag{33}$$

Wyższe wartości tego współczynnika są bardziej pożądane. Nie można go jednak wykorzystać do porównywania modeli o różnej liczbie stopni swobody z tego powodu, że model o większej złożoności będzie miał współczynnik pseudo R^2 większy w porównaniu z modelem prostszym.

Alternatywnie do oceny modelu można wykorzystać tzw. skorygowany współczynnik determinacji (*ang. adjusted pseudo* R^2), który wyraża się wzorem:

$$R^2_{Adj} = 1 - \frac{q - r_0}{q - r_1}(1 - R^2), \tag{34}$$

gdzie q to liczba komórek w tablicy kontyngencji (dla dwuwymiarowej tablicy 2×2 mamy $q = 4$) a r_0 i r_1 to liczba parametrów odpowiednio w przypadku modelu M_0 i M_1. Im większa jest wartość R^2_{Adj}, tym model M_1 jest lepiej dopasowany do danych. Współczynnik ten nie jest unormowany w przedziale $[0,1]$ i może przyjmować wartości ujemne.

TABELA 5. RÓWNANIE I SYMBOLICZNY ZAPIS MODELI LOG-LINIOWYCH DLA TRZECH ZMIENNYCH X, Y, Z

Nazwa modelu	Równanie modelu	Notacja nawiasowa
Model pełny	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$	[XYZ]
Homogeniczna zależność	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	[XY][XZ][YZ]
Warunkowa niezależność	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$	[XY][YZ]
	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$	[XY][XZ]
	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	[XZ][YZ]
Łączna niezależność	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	[XY][Z]
	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$	[XZ][Y]
	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$	[YZ][X]
Całkowita niezależność	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	[X][Y][Z]
Niezależność częściowa	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$	[XY]
	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_k^Z + \lambda_{ik}^{XZ}$	[XZ]
	$ln(m_{ijk}) = \mu + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$	[YZ]
Niezależność całkowita z 2 wpływami	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y$	[X][Y]
	$ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_k^Z$	[X][Z]
	$ln(m_{ijk}) = \mu + \lambda_j^Y + \lambda_k^Z$	[Y][Z]
Niezależność całkowita z 1 wpływem	$ln(m_{ijk}) = \mu + \lambda_i^X$	[X]
	$ln(m_{ijk}) = \mu + \lambda_j^Y$	[Y]
	$ln(m_{ijk}) = \mu + \lambda_k^Z$	[Z]
Model zerowy	$ln(m_{ijk}) = \mu$	[0]

3. Modele log-liniowe w estymacji populacji trudnych do zbadania

Metoda wielokrotnych połowów (*ang. capture-recapture*) jest często wykorzystywaną techniką estymacji populacji trudnych do zbadania, w której wykorzystuje się informacje z dwóch lub większej liczby źródeł danych, łączy się jednostki a następnie oszacowuje się liczbę jednostek, które nie występują w żadnym źródle. Przykładowo w przypadku dwóch źródeł danych A i B może mieć miejsce sytuacja, w której po połączeniu jednostek (deterministyczne lub probabilistyczne łączenie) występują jednostki tylko w źródle A a nie występują w źródle B, występują w źródle B i nie występują w źródle A oraz występują jednocześnie w źródle A i B. Zilustrowane to zostało w Tabeli 6.

TABELA 6. PRZYPADEK DWÓCH ŹRÓDEŁ – TABLICA KONTYNGENCJI 2x2

		Źródło B		∑
		Tak (1)	Nie (0)	
Źródło A	Tak (1)	n_{11}	n_{10}	n_{1+}
	Nie (0)	n_{01}	n_{00}	n_{0+}
∑		n_{+1}	n_{+2}	n

W powyższej tabeli Tak (1) oznacza, że dana jednostka występuje w odpowiednim źródle a Nie (0), że jednostka w tym źródle nie występuje. Przykładowo n_{01} oznacza liczbę jednostek, które nie występują w źródle A a występują w źródle B. Kluczową kwestią jest zatem oszacowanie liczebności n_{00} tj. liczby jednostek, które nie występują zarówno w źródle A jak i B. Ostatecznie oszacowaną liczebność populacji uzyskuje się poprzez dodanie wszystkich wartości z Tabeli 6 po wcześniejszym wyestymowaniu liczebności n_{00}.

Oszacowanie liczebności n_{00} może być uzyskane poprzez dopasowanie modelu log-liniowego do niekompletnej tablicy kontyngencji. Przykładowo w odniesieniu do Tabeli 6 wymiarów 2×2 odnoszących się do źródeł danych A i B pełen model log-liniowy [AB] może być przedstawiony w postaci (por. wzór 7):

$$ln(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \quad i,j = \{'\text{Tak}''\text{Nie}'\}, \tag{35}$$

gdzie m_{ij} oznacza oczekiwaną liczebność w komórce i,j. Ponieważ jednak komórka $m_{00} = m_{\text{Nie,Nie}}$ nie jest obserwowana, model [AB] ma jeden parametr za dużo i nie może być zatem estymowany. W takiej sytuacji można rozważyć model niezależności [A][B] postaci:

$$ln(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B, \tag{36}$$

który ma tylko trzy parametry do oszacowania w związku z brakiem efektu interakcji λ_{ij}^{AB}. W takiej sytuacji w związku z tym, że mamy trzy obserwowane komórki w Tabeli 6 oraz trzy parametry do oszacowania mamy w zasadzie do czynienia z modelem nasyconym. Po dopasowaniu tego modelu do danych możemy użyć oszacowanych parametrów do wyznaczenia liczebności brakującej komórki {'Nie', 'Nie'} a następnie wyznaczyć liczebność populacji poddanej analizie. Oszacowanie liczebności komórki m_{00} znajdujemy przy tym ze wzoru:

$$\hat{m}_{00} = \exp(\mu). \tag{37}$$

Podobne rozumowanie można przeprowadzić w odniesieniu do tablic trójdzielczych typu $2 \times 2 \times 2$ (por. Tabela 7) tj. w sytuacji gdy dysponujemy trzema źródłami danych A, B i C.

TABELA 7. PRZYPADEK TRZECH ŹRÓDEŁ – TABLICA KONTYNGENCJI 2x2x2

		Źródło C				Σ
		Tak (1)		Nie (0)		
		Źródło B		Źródło B		
		Tak (1)	Nie (0)	Tak (1)	Nie (0)	
Źródło A	Tak (1)	n_{111}	n_{101}	n_{110}	n_{100}	n_{1++}
	Nie (0)	n_{011}	n_{001}	n_{010}	n_{000}	n_{0++}
Σ		n_{+11}	n_{+01}	n_{+10}	n_{+00}	n

Tabela 7 może przedstawiać sytuację trzech źródeł, na przykład trzech rejestrów administracyjnych, dwóch rejestrów administracyjnych i badania reprezentacyjnego czy spisu. Podobnie jak w przypadku tabeli 2×2 istotne jest określenie przynależności do poszczególnego źródła (oznaczone jako Tak/Nie). Również i w tym przypadku chcemy oszacować to czego nie możemy odczytać z tabeli tj. n_{000}. Na potrzeby estymacji liczebności n_{000} można również wykorzystać koncepcję modeli log-liniowych. W tym celu budujemy model log-liniowy postaci (bez efektu głównego λ_{ijk}^{ABC}) :

$$ln(m_{ijk}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}, \quad (38)$$

który musimy ograniczyć przez:

$$\lambda_0^A = \lambda_0^B = \lambda_0^C = \lambda_{00}^{AB} = \lambda_{10}^{AB} = \lambda_{01}^{AB} = \lambda_{00}^{AC} = \lambda_{10}^{AC} = \lambda_{01}^{AC} = \lambda_{00}^{BC} = \lambda_{10}^{BC} = \lambda_{01}^{BC} = 0,$$

,

aby móc oszacować parametry. Dodatkowym założeniem jest to, że nie występuje interakcja między A, B i C, tj. $\lambda_{ijk}^{ABC} = 0$. Model ten oznacza się przez $[AB][BC][AC]$ - por. Tabela 5. Oszacowanie brakującej liczby jednostek populacji otrzymujemy ze wzoru:

$$\hat{m}_{000} = \exp(\mu). \quad (39)$$

Jak to zostało wspomniane w raporcie pośrednim pracy badawczej w przypadku estymacji wielkości populacji możliwe jest wykorzystanie zmiennych pomocniczych, którymi mogą być przykładowo płeć, grupy wieku czy województwa. Celem jest z jednej strony obejście jednego z założeń metody *capture-recapture* (o stałej stopie pokrycia przez źródło w populacji – *enumerate rate*) i uwzględnienie faktu heterogeniczności przynależności poszczególnych jednostek do źródeł. Wykorzystanie zmiennych pomocniczych w kontekście modeli log-liniowych rozważa m.in. Gerritse (2016), Coumans i inn. (2017), Peter i inn. (2012) czy Zwane i van der Heijden (2016). Wyróżniamy przy tym dwa podejścia, które determinowane są dostępnością zmiennych we wszystkich, niektórych lub tylko w jednym źródle. Pierwsze podejście określa się w literaturze jako *fully observed covariates*, a drugie *partially observed covariates*. W obydwu przypadkach można wykorzystać modele log-liniowe do oszacowania poszczególnych elementów populacji. Tego typu podejście zostało również zastosowane w niniejszej pracy badawczej. Przykładowo w przypadku dwuwymiarowej tabeli kontyngencji 2×2 oprócz przynależności do dwóch źródeł A i B można rozpatrywać dodatkową cechę X (na przykład płeć) przez co należy rozszerzyć tabelę do trójdzielczej Tabeli 8 oraz dopasować model log-liniowy [AX][BX] postaci:

$$ln(m_{ijx}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX}, \quad (40)$$

gdzie λ_{ix}^{AX} oraz λ_{jx}^{BX} oznaczają efekty interakcji pomiędzy zmienną pomocniczą X i źródłami danych A oraz

B odpowiednio.

TABELA 8. PRZYPADEK DWÓCH ŹRÓDEŁ A I B ORAZ JEDNEJ ZMIENNEJ POMOCNICZEJ X

		Zmienna X				Σ
		X_1		X_2		
		Źródło B		Źródło B		
		Tak (1)	Nie (0)	Tak (1)	Nie (0)	
Źródło A	Tak (1)	n_{111}	n_{101}	n_{110}	n_{100}	n_{1++}
	Nie (0)	n_{011}	n_{001}	n_{010}	n_{000}	n_{0++}
Σ		n_{+11}	n_{+01}	n_{+10}	n_{+00}	n

W przypadku dwóch źródeł A i B oraz jednej zmiennej pomocniczej X, przyjmującej przykładowo dwa warianty X_1 oraz X_2, (np. mężczyzna i kobieta), mamy do czynienia z trójdzielczą tablicą kontyngencji $2 \times 2 \times 2$, w której jednak brakujące liczebności, które należy oszacować to n_{001} oraz n_{000}. Mamy zatem 6 komórek, dla których znane są obserwowane liczebności w Tabeli 8 w związku z czym model (40) zawiera sześć parametrów, które należy oszacować (nasycony model log-liniowy). Po dopasowaniu modelu do danych brakujące liczebności komórek znajdujemy ze wzorów: $\widehat{m}_{000} = \exp(\mu)$ oraz $\widehat{m}_{001} = \exp\left(\mu + \lambda_{X_1}^{X}\right)$. Powyższe rozumowanie w naturalny sposób można rozszerzyć na większą liczbę zmiennych pomocniczych oraz liczbę analizowanych źródeł. Zwiększa się przez to w oczywisty sposób złożoność analizowanych modeli log-liniowych, jednak wykorzystanie odpowiednich pakietów języka R znacznie skraca proces estymacji wszystkich możliwych do zbudowania modeli.

Literatura

Brzezińska J. (2015), *Analiza logarytmiczno-liniowa. Teoria i zastosowania z wykorzystaniem programu R*, Wydawnictwo C.H. Beck, Warszawa.

Coumans A.M. Cruyff M., Van der Heijden., Wolf J., Schmeets H. (2017), *Estimating homelessness in the Netherlands using a capture-recapture approach*, Social Indicators Research, 130(1):189–212.

Gerritse S.Ch. (2016), *An application of population size estimation to official statistics: Sensitivity of model assumptions and the effect of implied coverage*, PhD thesis, Utrecht University.

Goodman L.A. (1964), *Simple Methods for Analyzing Three-factor Interaction in Contingency Tables*, Journal of the American Statistical Association, 59, 319-352.

Goodman L.A. (1968), *The Analysis of Cross-Classified Data: Independence, Quasi-independence, and Interactions in Contingency Tables with or without Missing Values*, Journal of the American Statistical Association, 63, 1091–1131.

Goodman L.A. (1969), *On Partitioning χ^2 and Detecting Partial Association in Three-way Contingency Tables*, Journal of the Royal Statistical Society. Series B (Methodological), 486-498.

Peter G.M., Van Der Heijden., Whittaker J., Cruyff M., Bakker B., Van der Vliet R. (2012), *People Born in the Middle East but Residing in the Netherlands: Invariant Population Size Estimates and the Role of Active and Passive Covariates*, The Annals of Applied Statistics, Vol. 6, No. 3, 831-852.

Stokes M.E., Davis C.S., Koch G.G. (2012), *Categorical Data Analysis Using SAS, Third Edition*, SAS Institute

Inc., Cary, NC, USA.

Zwane E., van der Heijden P.G.M. (2005), *Population estimation using the multiple system estimator in the presence of continuous covariates*, Statistical Modelling, 5(1): 39–52.