PRZEGLĄD PSYCHOLOGICZNY, 2007, TOM 50, Nr 1, 25-44

Modele log-liniowe i ich zastosowania w psychologii

Małgorzata Półtorak*
Instytut Studiów Społecznych Uniwersytetu Warszawskiego

LOG-LINEAR MODELS AND THEIR APPLICATIONS IN PSYCHOLOGY

Abstract. Log-linear analysis is a technique for the multivariate analysis of categoric or nominal scale data. Many psychologists are often confused by various "less familiar" techniques and thus unlikely to use it. The aim of this article is to help them to get to know the log-linear models, make clear their basic terminology and give a real example of their application in psychology.

Jedną z podstawowych metod analizy danych jest badanie tabel krzyżowych, lub inaczej tabel kontyngencyjnych. Na przykład w psychologii klinicznej mogą to być tabele liczebności występowania symptomów zaburzeń u pacjentów różnej płci, w różnych grupach wiekowych i o różnych cechach osobowościowych, a w psychologii ekonomicznej tabele preferencji konsumenckich ze względu na rodzaj produktu, wiek, płeć, temperament itd.

Analiza log-liniowa (czyli logarytmiczno-liniowa) to jeden z bardziej wyrafinowanych sposobów patrzenia na tabele krzyżowe, który w szczególności pozwala testować istotność statystyczną wpływu różnych czynników ujętych w tabeli i ich interakcji. W ciągu ostatnich lat, odkąd do wykonywania skomplikowanych obliczeń powszechnie używa się komputerów, ten rodzaj analizy zyskuje coraz większość popularność, zwłaszcza w badaniach problemów nauk

* Serdecznie dziękuję Pani Profesor Elżbiecie Aranowskiej, pod której kierunkiem powstały moje prace roczne (Półtorak, 2001a; 2001b) będące podstawą niniejszego artykułu. Adres do korespondencji: Instytut Nauk Społecznych Uniwersytetu Warszawskiego, ul. Stawki 5/7, 00-183 Warszawa; e-mail: mpoltorak@uw.edu.pl

społecznych, gdzie zmiennych często nie da się przedstawić inaczej niż na skali nominalnej.

Mimo iż modele log-liniowe często wydają się psychologom czymś zupełnie nowym i nieznanym, bliższe przyjrzenie się im pozwala dostrzec wiele podobieństw do zwykłej regresji liniowej. Metoda regresji jest zwykle stosowana, jeśli zmienna zależna jest określona na skali przedziałowej lub stosunkowej. Jednakże gdy zmienna zależna ma rozkład dwupunktowy, wtedy wpływ zmiennych niezależnych może być interpretowany jako oddziaływanie na prawdopodobieństwo (o wartościach od 0 do 1), że zmienna zależna przybierze pożądaną wartość. W analizie log-liniowej jest analogicznie, z tą różnicą, że przedmiotem analizy jest wpływ zmiennych niezależnych na stosunek szans (o wartościach od 0 do ∞) należenia do danej kategorii (to znaczy stosunek liczby przypadków należących do tej kategorii do liczby przypadków nie należących do niej).

Niniejszy artykuł rozpoczyna omówienie podstawowych pojęć związanych z analizą log-liniową, następnie zaś przedstawione są różne rodzaje modeli oraz niektóre bardziej szczegółowe właściwości metody i techniki w niej używane. W celu jak najlepszego zilustrowania metody, w ostatniej części artykułu zamieszczony został konkretny, rzeczywisty przykład analizy danych z badań przeprowadzonych przez autorkę. Do obliczeń numerycznych użyto programu STATISTICA.

PODSTAWOWE POJĘCIA I ICH INTERPRETACJA

Tabela kontyngencji. Tabela kontyngencji to tabela liczebności przypadków w poszczególnych kategoriach zmiennych (inaczej: tabela rozkładu frekwencji). Przykładem tabeli krzyżowej 2×2 jest tabela 1 zawierająca liczebności w badanej próbie osób cierpiących i nie cierpiących na depresję oraz osób popełniających samobójstwo i nie popełniających go. Przez pełną analogię można wyobrazić sobie wielozmiennową tabelę kontyngencji.

Tabela 1.

Przykład dwuzmiennowej (dwuwymiarowej) tabeli kontyngencji

		Depresja (D)	
		+	–
Popełnianie samobójstw (S)	+	$f_{11} = 90$	$f_{12} = 10$
	–	$f_{21} = 10$	$f_{22} = 990$

Model. Słowa „model" używam w następującym sensie: Model jest określeniem oczekiwanych frekwencji (F_{ij}) w komórkach tabeli o dowolnej skończonej

liczbie wierszy i kolumn jako funkcji parametrów reprezentujących wpływ poszczególnych kategorii zmiennych nominalnych lub porządkowych (tzw. czynników) zawartych w tabeli i ich wzajemnych związków. Mówimy, że model „pasuje" do danych (a więc dobrze je wyjaśnia), jeśli frekwencje oczekiwane dobrze aproksymują frekwencje zaobserwowane, czyli mało się od nich różnią.

Szansa *i*-tej kategorii zmiennej. Szansą *i*-tej kategorii zmiennej (szansą, że przypadek należy do *i*-tej kategorii danej zmiennej) nazywam stosunek liczby przypadków w kategorii, dla której określamy szanse (*i*-tej), do liczby przypadków w pozostałych kategoriach (1, 2, ..., $i - 1$, $i + 1$, ...), zatem jest to dowolna nieujemna liczba rzeczywista, która może być, ale nie musi, mniejsza od jedności, stąd nie należy mylić pojęcia szansy z pojęciem prawdopodobieństwa. Matematycznie jest to równoznaczne stosunkowi prawdopodobieństwa, że przypadek należy do tej kategorii do prawdopodobieństwa, że przypadek nie należy do tej kategorii. Na przykład jeśli prawdopodobieństwo należenia jednostki do grupy osób cierpiących na depresję wynosi 100/1100 = 1/11 (a więc prawdopodobieństwo nienależenia jednostki do tej grupy wynosi 1000/1100 = = 10/11), to możemy powiedzieć, że szansa na to, że osoba cierpi na depresję, wynosi „jeden do dziesięciu" ((1/11)/(10/11) = 1/10 = 1:10), a szansa, że osoba nie ma depresji – 10:1.

Szansa brzegowa *i*-tej kategorii zmiennej. Szansa brzegowa (zaobserwowana lub oczekiwana) *i*-tej kategorii zmiennej jest to szansa określana bez brania pod uwagę innych zmiennych, tzn. stosunek zsumowanych frekwencji (zaobserwowanych lub oczekiwanych) tej zmiennej. Na przykładzie tabeli 1, szansa brzegowa popełnienia samobójstwa (brzegowa, czyli niezależna od depresji) wynosi

$$\frac{90+10}{10+990} = \frac{100}{1000} = \frac{1}{10} = 1{:}10.$$

Szansa warunkowa *i*-tej kategorii zmiennej. Szansa warunkowa jest to szansa określana dla *i*-tej kategorii danej zmiennej przy ustalonej kategorii innej zmiennej, tzn. stosunek liczebności (frekwencji) przypadków należących do żądanej (*i*-tej) kategorii pierwszej zmiennej i pewnej ustalonej kategorii drugiej zmiennej do liczebności (frekwencji) przypadków nie należących do żądanej kategorii pierwszej zmiennej i należących do tej samej kategorii drugiej zmiennej, np. dla tabeli 1 warunkowa szansa popełnienia samobójstwa pod warunkiem bycia w grupie osób cierpiących na depresję wynosi 90/10 = 9:1.

Zmienne nieskorelowane. Dwie zmienne będziemy nazywać nieskorelowanymi, jeśli wszystkie szanse warunkowe jednej zmiennej względem drugiej (i na odwrót) są równe lub prawie równe i tym samym bliskie szansie brzegowej tej zmiennej (w przypadku depresji (D) i samobójstwa (S) byłoby tak, gdy-

by szanse popełnienia samobójstwa wśród osób depresyjnych i nie depresyjnych były równe oraz gdyby szanse cierpienia na depresję wśród osób popełniających i nie popełniających samobójstw były równe).

Proporcja szans. Proporcja szans (SD) jest statystyką używaną do bezpośredniego porównywania szans warunkowych, wykorzystywaną w modelach log-liniowych. Na przykładzie tabeli 1 proporcja szans zaobserwowanych to:

$$\mathrm{SD} = \frac{f_{11}/f_{21}}{f_{12}/f_{22}} = \frac{f_{11}f_{22}}{f_{21}f_{12}} = \frac{90\times 990}{10\times 10} = 891.$$

Łatwo zauważyć, że proporcja szans jest zawsze nieujemna i może osiągać dowolnie duże wartości. Jeśli proporcja szans jest równa 1, znaczy to, że porównywane szanse warunkowe są identyczne i między zmiennymi nie ma żadnego związku. Jeśli jest większa od 1 (tak jest w powyższym przykładzie), to zmienne są skorelowane dodatnio, a jeśli mniejsza od 1, to ujemnie. Znak korelacji jest oczywiście ustalony arbitralnie, jako że mamy do czynienia ze skalą nominalną.

OPIS METODY

Uogólnione modele log-liniowe

Istnieją dwa rodzaje modeli log-liniowych. Pierwszy z nich to uogólniony model log-liniowy, w którym nie rozróżnia się zmiennych zależnych i niezależnych, a frekwencje oczekiwane są analizowane jako funkcje wszystkich zmiennych występujących w modelu. W drugim modelu, zwanym modelem logitowym, co najmniej jedna ze zmiennych określana jest jako zmienna zależna. Analizowane są wtedy oczekiwane szanse tej zmiennej (Ω_{ij}) (*omega*), traktowanej jako funkcja zmiennych niezależnych. Model logitowy przypomina zwykłą regresję liniową.

Modele nasycone

W modelu nasyconym występują efekty wszystkich możliwych czynników. Dla tabeli czteropolowej, takiej jak tabela 2, ma on postać:

$$F_{ij} = \eta \tau_i^A \tau_j^B \tau_{ij}^{AB},$$

gdzie F_{ij} oznacza oczekiwaną frekwencję przypadków w komórce o indeksie (i, j), η (*eta*) jest średnią geometryczną liczby przypadków w każdej komórce, zaś poszczególne τ (*tau*) oznaczają efekty, jakie zmienne wywierają na frekwencje w odpowiednich komórkach (indeksy dolne oznaczają numer kategorii, a indeksy górne to symbole poszczególnych zmiennych).

Tabela 2.

Tabela dwuzmiennowa frekwencji zaobserwowanych

		Czynnik B	
		+	–
Czynnik A	+	f_{11}	f_{12}
	–	f_{21}	f_{22}

Parametry wpływu (τ), czyli efekty oddziaływania poszczególnych czynników, są związane z szansami i ich proporcjami[1]. Efekt τ_i^A (efekt i-tej kategorii zmiennej A), jeden dla każdej kategorii zmiennej A, występuje, jeśli iloczyn szans warunkowych tej kategorii zmiennej A jest różny od jedności, tzn. $(F_{11}/F_{21})\times(F_{12}/F_{22}) \neq 1$. Efekt τ_j^B (efekt j-tej kategorii zmiennej B), jeden dla każdej kategorii zmiennej B, występuje, jeśli iloczyn szans warunkowych tej kategorii zmiennej B jest różny od jedności, tzn. $(F_{11}/F_{12}) \times (F_{21}/F_{22}) \neq 1$. Efekty τ_{ij}^{AB} występują, jeśli A i B są skorelowane (czyli proporcje wszystkich szans warunkowych są różne od 1).

Jeżeli dany czynnik (lub interakcja czynników) nie wywiera wpływu na frekwencje, to odpowiadające jego kategoriom parametry τ są równe 1, gdyż wówczas wszystkie związane z nim szanse są sobie równe. Za pomocą powyższych parametrów frekwencje oczekiwane dla powyższego modelu można przedstawić tak jak w tabeli 3.

Tabela 3

Tabela dwuzmiennowa frekwencji oczekiwanych

		Czynnik B	
		+	–
Czynnik A	+	$F_{11} = \eta\tau_1^{\mathrm{A}}\ \tau_1^{\mathrm{B}}\ \tau_{11}^{\mathrm{AB}}$	$F_{12} = \eta\tau_1^{\mathrm{A}}\ \tau_2^{\mathrm{B}}\ \tau_{12}^{\mathrm{AB}}$
	–	$F_{21} = \eta\tau_2^{\mathrm{A}}\ \tau_1^{\mathrm{B}}\ \tau_{21}^{\mathrm{AB}}$	$F_{22} = \eta\tau_2^{\mathrm{A}}\ \tau_2^{\mathrm{B}}\ \tau_{22}^{\mathrm{AB}}$

Dla zmiennych o rozkładach dwupunktowych, takich jak A i B w powyższych tabelach, parametry wpływu τ dla wszystkich kategorii zmiennych są wzajemnie odwrotne (tzn. $\tau_1^A\ \tau_2^A = 1$, $\tau_1^B\ \tau_2^B = 1$). Mamy więc:

$$\tau^A = \tau_1^A = 1/\tau_2^A,$$

$$\tau^B = \tau_1^B = 1/\tau_2^B.$$

[1] Zob. punkt Podstawowe pojęcia i ich interpretacja.

Podobnie iloczyn τ^{AB} w każdym wierszu i w każdej kolumnie tabeli jest równy 1:

$$\tau^{AB} = \tau_{11}^{AB} = \tau_{22}^{AB} = 1/\tau_{12}^{AB} = 1/\tau_{21}^{AB}.$$

Ponieważ w modelu nasyconym występuje więcej parametrów niż frekwencji komórek w tabeli, model ten nie mógłby być estymowany bez powyższych ograniczeń. Te ograniczenia znaczą, że tylko cztery parametry z danych dziewięciu są niezależne (η, jedno z τ_i^A, jedno z τ_j^B i jedno z τ_{ij}^{AB}). Ponieważ mamy cztery parametry niezależne i cztery komórki w tabeli, otrzymujemy model perfekcyjnie odzwierciedlający frekwencje zaobserwowane, w związku z czym możemy traktować je jako identyczne z oczekiwanymi.

Wykorzystując równania z tabeli 3, możemy wyprowadzić wzory określające zależności parametrów τ i η od frekwencji oczekiwanych. Po przekształceniach otrzymujemy:

$\tau^{AB} = (F_{11}F_{22} / F_{21}F_{12})^{1/4}$ (co oznacza, że efekt posiadania czynnika A i czynnika B, podniesiony do potęgi czwartej, jest równy proporcji szans warunkowych dowolnej ze zmiennych A i B),

$\tau^{A} = (F_{11}F_{12} / F_{21}F_{22})^{1/4}$ (co oznacza, że efekt posiadania czynnika A, podniesiony do potęgi czwartej, jest równy iloczynowi szans warunkowych zmiennej A),

$\tau^{B} = (F_{11}F_{21} / F_{12}F_{22})^{1/4}$ (co oznacza, że efekt posiadania czynnika B, podniesiony do potęgi czwartej, jest równy iloczynowi szans warunkowych zmiennej B),

$\eta = (F_{11}F_{21}F_{12}F_{22})^{1/4}$ (co oznacza średnią geometryczną frekwencji oczekiwanych).

Modele nienasycone

Ponieważ modele nasycone reprezentują oczekiwane frekwencje komórkowe jako funkcje wszystkich możliwych efektów (i parametru *eta* – η), nie są one modelami „oszczędnymi". Mniej skomplikowane modele można konstruować zakładając, że część zmiennych nie wywiera wpływu na frekwencje, czyli przyrównując wszystkie parametry wpływu ich kategorii (τ) do jedynki (analogicznie jak w równaniu regresji można zakładać, że część współczynników jest równa zero). W takich „nienasyconych" modelach frekwencje oczekiwane są zawsze mniej lub bardziej rozbieżne z obserwowanymi. Dla przykładu przedstawię pewne modele nienasycone dla danych z tabeli 2:

1. Model, który zakłada, że czynniki A i B są nieskorelowane (w sensie niezależności mierzonej testem χ^2), czyli taki, gdzie $\tau^{AB} = 1$:

$$F_{ij} = \eta\tau_i^A \tau_j^B.$$

2. Model, w którym dodatkowo czynnik B nie ma wpływu na frekwencje ($\tau^B = 1$):

$$F_{ij} = \eta\tau_i^A .$$

3. Model, w którym wszystkie τ są równe 1:

$$F_{ij} = \eta.$$

Jako przykład danych, dla których pewnie warto byłoby stosować modele nienasycone zamiast nasyconych, można wyobrazić sobie sytuację, kiedy mielibyśmy dane dotyczące liczby popełnianych samobójstw wśród osób różnej płci. Moglibyśmy na przykład założyć, że płeć i popełnianie samobójstw są zmiennymi niezależnymi i rozważać model postaci takiej, jak w punkcie 1 powyżej. W przypadku, gdyby próba była dodatkowo zrównoważona pod względem płci, bardziej użyteczne mogłoby się okazać rozważanie modelu zakładającego również, że płeć nie ma wpływu na frekwencje w poszczególnych grupach (postaci takiej, jak w punkcie 2).

Notacja i wzór Goodmana

Wszystkie modele przedstawione dotychczas zostały zaprezentowane w formie multiplikatywnej (to znaczy takiej, że frekwencje są iloczynami parametrów). Łatwo zauważyć, że poprzez zlogarytmowanie obu stron[3] można przekształcić te równania do postaci liniowej względem logarytmów zmiennych (tzw. logarytmiczno-addytywnej). Równania takie są nazywane log-liniowymi (i stąd też nazwa metody). Formy multiplikatywna i logarytmiczno-addytywna równań są matematycznie tożsame.

Postacią log-liniową równania na frekwencje z modelu nasyconego $F_{ij} = \eta\tau_i^A \tau_j^B \tau_{ij}^{AB}$ jest:

$$\ln(F_{ij}) = \ln(\eta\tau_i^A \tau_j^B \tau_{ij}^{AB}) = \ln(\eta) + \ln(\tau_i^A) + \ln(\tau_j^B) + \ln(\tau_{ij}^{AB})$$

lub inaczej, według tzw. notacji Goodmana (1972),

$$G_{ij} = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB},$$

gdzie poszczególne λ (*lambda*) są logarytmami naturalnymi odpowiednich τ, θ (*teta*) jest logarytmem naturalnym η, a G_{ij} to logarytm naturalny F_{ij}. Log-liniowe wersje równań upodabniają je do zwykłej regresji liniowej. Logarytm naturalny z jedynki wynosi zero, zatem brak pewnych *lambd* w modelu ozna-

[3] Mam tu na myśli logarytm naturalny.

cza, iż wartość tych parametrów wynosi 0, co jest równoznaczne z brakiem odpowiadających im τ w postaci multiplikatywnej równania.
Statystyczne znaczenie parametrów wpływu w modelu nasyconym można łatwo wyznaczyć za pomocą log-liniowej formy równania, a konkretnie wzoru Goodmana (1972) na odchylenie standardowe *lambd*:

$$\hat{S}_\lambda = \sqrt{\frac{\sum_i \sum_j (1/f_{ij})}{c^2}},$$

gdzie *c* oznacza liczbę komórek tabeli.

Jeśli wartość oczekiwana *lambdy* jest równa 0 (a więc oczekiwana wartość *tau* wynosi 1 i dana zmienna nie wywiera wpływu na frekwencje), rozkład wystandaryzowanej $\hat{\lambda} = \lambda / S_\lambda$ ($\hat{\lambda}$ oznacza, ile *lambd* mieści się na odcinku S_λ jednego odchylenia standardowego) dla dużych prób przybliża się rozkładem normalnym o zerowej wartości oczekiwanej i jednostkowej wariancji. A zatem, tak jak w zwykłej regresji liniowej, wartość $|\hat{\lambda}|$ większa niż 1,96 byłaby znacząca na poziomie istotności 0,05.

Mimo iż powyższą standaryzacja *lambd* odnosi się ściśle tylko do modeli nasyconych, wartość S_λ wyznacza górną granicę statystyki testu przy testach istotności parametrów modeli nienasyconych.

Zgodność zaobserwowanych i oczekiwanych brzegowych rozkładów frekwencji w modelach hierarchicznych

W dalszej części ograniczę typy omawianych modeli do tak zwanych modeli hierarchicznych, czyli takich, w których uwzględniając efekt interakcji kilku czynników, bierzemy też pod uwagę wszystkie możliwe efekty interakcji poszczególnych z tych czynników i efekty pojedynczych czynników. Dla uproszczenia notacji do oznaczania modeli hierarchicznych używa się liter odpowiadających parametrom skorelowanym, wziętych w nawiasy klamrowe. Każdy nawias zawiera parametry wpływu najwyższego rzędu uwzględnione w modelu (wszystkie one są różne od 1 w wersji multiplikatywnej i od 0 w addytywnej). Na przykład, jeśli mamy czynniki A, B, C i D, tworzące czterowymiarową tabelę kontyngencji, to model hierarchiczny nasycony {ABCD} zawiera parametry wpływu

$$\tau^{ABCD}, \tau^{ABC}, \tau^{ABD}, \tau^{ACD}, \tau^{BCD}, \tau^{AB}, \tau^{AC}, \tau^{AD}, \tau^{BC}, \tau^{BD}, \tau^{CD}, \tau^{A}, \tau^{B}, \tau^{C}, \tau^{D},$$

(no i oczywiście parametr η). Zapis {ABCD} jest równoznaczny zapisowi $F_{ijkl} = \eta \tau^{ABCD}_{ijkl} \tau^{ABC}_{ijk} \tau^{ABD}_{ijl} \tau^{ACD}_{ikl} \tau^{BCD}_{jkl} \tau^{AB}_{ij} \tau^{AC}_{ik} \tau^{AD}_{il} \tau^{BC}_{jk} \tau^{BD}_{jl} \tau^{CD}_{kl} \tau^{A}_{i} \tau^{B}_{j} \tau^{C}_{k} \tau^{D}_{l}$ dla każdych i, j, k, l; model {ABC}, jeden z „podmodeli" modelu nasyconego, zawiera parametry wpływu $\tau^{ABC}, \tau^{AC}, \tau^{AB}, \tau^{BC}, \tau^{A}, \tau^{B}, \tau^{C}$ oraz η i jest równoznacz-

ny zapisowi $F_{ijkl} = \eta\tau_{ijk}^{ABC}\tau_{ij}^{AB}\tau_{ik}^{AC}\tau_{jk}^{BC}\tau_{i}^{A}\tau_{j}^{B}\tau_{k}^{C}$ dla każdych i, j, k; model {C}{D} zawiera parametry τ^{C}, τ^{D} oraz η i jest równoznaczny zapisowi $F_{ij} = \eta\tau_{i}^{C}\tau_{j}^{D}$ dla każdych i, j; model {AB}{C} zawiera parametry τ^{AB}, τ^{A}, τ^{B}, τ^{C} oraz η i jest równoznaczny zapisowi $F_{ijkl} = \eta\tau_{ij}^{AB}\tau_{i}^{A}\tau_{j}^{B}\tau_{k}^{C}$ dla każdych i, j, k, zaś model {ABC}{CD} zawiera parametry τ^{ABC}, τ^{AC}, τ^{AB}, τ^{BC}, τ^{CD}, τ^{A}, τ^{B}, τ^{C}, τ^{D} oraz η i jest równoznaczny zapisowi $F_{ijkl} = \eta\tau_{ijk}^{ABC}\tau_{ij}^{AB}\tau_{ik}^{AC}\tau_{jk}^{BC}\tau_{kl}^{CD}\tau_{i}^{A}\tau_{j}^{B}\tau_{k}^{C}\tau_{l}^{D}$ dla każdych i, j, k, l.

Ważną właściwością hierarchicznych modeli log-liniowych jest całkowita zgodność zaobserwowanych i oczekiwanych brzegowych rozkładów frekwencji. „Podtabele" frekwencji są wydzielane z pełnej tabeli przez zmienne pogrupowane w nawiasach klamrowych (przez pojedyncze zmienne i wszystkie wchodzące w skład modelu interakcje). Estymowane oczekiwane frekwencje dla konkretnych poziomów wybranych czynników i ich interakcji (jeśli model je uwzględnia) muszą być równe frekwencjom zaobserwowanym w odpowiednich „podtabelach". Oznacza to, że frekwencje oczekiwane zawsze „pasują" do danych dla wyszczególnionych rozkładów brzegowych.

W przypadku modelu {A}{B} (tzn. takiego, że $F_{ij} = \eta\tau_{i}^{A}\tau_{j}^{B}$) dla danych z tabeli 2 znaczy to, że rozkłady frekwencji oczekiwanych pojedynczych zmiennych A, B są odpowiednio identyczne z sumami wierszy i kolumn frekwencji zaobserwowanych, a więc:

$$f_{11} + f_{21} = F_{11} + F_{21},$$
$$f_{12} + f_{22} = F_{12} + F_{22},$$
$$f_{11} + f_{12} = F_{11} + F_{12},$$
$$f_{21} + f_{22} = F_{21} + F_{22}.$$

Mimo iż frekwencje oczekiwane różnią się od obserwowanych w obrębie wierszy i kolumn (na przykład może być $f_{11} \neq F_{11}$), dla rozkładów brzegowych zmiennych A i B w pełni odpowiadają one danym. Proporcje szans oczekiwanych dla różnych poziomów czynników są równe 1, co potwierdza zakładany brak korelacji zmiennych.

Dobieranie frekwencji oczekiwanych

W modelach wielozmiennowych frekwencje dla poszczególnych konfiguracji poziomów czynników nie są, tak jak w modelach dwuzmiennowych, prostą sumą wierszy lub kolumn tabeli, a do ich wyznaczania potrzebne są złożone

algorytmy. Dwa najbardziej znane to algorytm Deminga-Stephena, zwany algorytmem „iteracyjnego dopasowywania proporcjonalnego", oraz algorytm Newtona-Raphsona (pierwszy z nich jest wykorzystywany na przykład w programie „Statistica" – por. StatSoft, Inc., 1995), drugi zaś w programie SPSS (por. Podręcznik użytkownika dołączony do programu). Mimo iż ten ostatni jest ogólniejszy, ograniczę się tylko do pobieżnego omówienia prostszego i powszechniej używanego algorytmu iteracyjnego dopasowywania proporcjonalnego.
W procedurze estymacji wykorzystywana jest (omówiona wcześniej) własność zgodności zaobserwowanych i oczekiwanych brzegowych rozkładów frekwencji. Wszystkie parametry nie uwzględnione w modelu mają przypisaną wartość 1. Podczas całego procesu estymacji powstają estymatory największej wiarygodności oczekiwanych frekwencji komórkowych, które są sukcesywnie dopasowywane do wszystkich rozkładów brzegowych zmiennych wyszczególnionych przez model (standardowo początkowe wartości we wszystkich komórkach są równe 1), na przykład w modelu {AB}{ACD}{BCD} „dopasowana" zostaje najpierw tablica odpowiadająca {AB}, potem {ACD}, a na końcu {BCD}. Proces dopasowywania jest kontynuowany dopóty, dopóki różnice między ostatnio uzyskanymi a poprzednimi wynikami estymacji nie są wystarczająco małe. Dokończeniem całej procedury jest wyznaczenie estymatorów parametrów wpływu dla poszczególnych czynników i ich interakcji.

Modele logitowe

Poniżej jest opisana druga forma modeli log-liniowych, czyli modele logitowe. W takich modelach zakłada się, że jedna wybrana zmienna jest zmienną zależną (objaśnianą), pozostałe zaś niezależnymi (objaśniającymi). Przedmiotem analizy są proporcje frekwencji oczekiwanych zmiennej zależnej, a właściwie ich logarytmów naturalnych. Aby porównać uogólniony model log-liniowy z logitowym, wyobraźmy sobie 3-czynnikowy model nasycony, gdzie czynnik A (o rozkładzie dwupunktowym) będziemy traktować jako zmienną zależną.
Wiemy, że w modelach log-liniowych frekwencje oczekiwane są funkcjami różnych parametrów wpływu, czyli że dla dowolnych *i*, *j*, *k* (gdzie *i*, *j*, *k* to odpowiednio indeks kategorii czynnika A, B i C) zachodzi:

$$F_{ijkl} = \eta \tau_i^A \tau_j^B \tau_k^C \tau_{ij}^{AB} \tau_{ik}^{AC} \tau_{jk}^{BC} \tau_{ijk}^{ABC} .$$

Na tej podstawie łatwo możemy wyprowadzić wzory na szanse oczekiwane kategorii czynnika A, na przykład dla pierwszej kategorii dla dowolnych *j* oraz *k* mamy:

$$\frac{F_{1jk}}{F_{2jk}} = \frac{\eta \tau_1^A \tau_{1j}^{AB} \tau_{1k}^{AC} \tau_{1jk}^{ABC}}{\eta \tau_2^A \tau_{2j}^{AB} \tau_{2k}^{AC} \tau_{2jk}^{ABC}} = (\tau^A)^2 (\tau_j^{AB})^2 (\tau_k^{AC})^2 (\tau^{ABC}_{jk})^2$$

(gdyż $\tau_1^A = 1/\tau_2^A$, $\tau_j^{AB} = \tau_{1j}^{AB} = 1/\tau_{2j}^{AB}$ itd.),

zaś po zlogarytmowaniu :

$$\ln(F_{1jk}/F_{2jk}) = 2\ln(\tau^A) + 2\ln(\tau_j^{AB}) + 2\ln(\tau_k^{AC}) + 2\ln(\tau_{jk}^{ABC}) = 2\lambda^A + 2\lambda_j^{AB} + 2\lambda_k^{AC} + 2\lambda_{jk}^{ABC},$$

gdzie *lambdy* są logarytmami naturalnymi z *tau*. W notacji Goodmana powyższe równanie ma postać:

$$\Phi_{jk}^{A} = \beta^{A} + \beta_{j}^{AB} + \beta_{k}^{AC} + \beta_{jk}^{ABC},$$

gdzie Φ_{jk}^{A} (*fi*) jest logarytmem naturalnym brzegowej szansy oczekiwanej pierwszej kategorii czynnika A, a każda β (*beta*) – to podwojona odpowiadająca jej λ.

W celu dokładniejszego zilustrowania różnic między uogólnionym modelem log-liniowym a logitowym rozważmy przykład czterech czynników. Wyobraźmy sobie, że mamy czterozmiennową tabelę kontyngencji ze zmiennymi A, B, C, D (zmienne: A, B, D – dychotomiczne, C – o trzech kategoriach). Dla przykładu: niech A oznacza niski vs wysoki poziom stresu w pracy, B – wewnętrzne vs zewnętrzne umiejscowienie kontroli, C – niski, średni lub wysoki poziom wsparcia społecznego, a D – poczucie samoskuteczności vs jego brak. W naszym modelu będziemy przyjmować, że szanse oczekiwane kategorii czynnika A (poziomu stresu w pracy) zależą od czynników B, C, D (umiejscowienia kontroli, poziomu wsparcia społecznego i poczucia samoskuteczności) oraz interakcji czynników C i D (poziomu wsparcia oraz poczucia samoskuteczności). Równanie logitowe dla tego modelu to:

$$\Phi_{ijk}^{A} = \beta^{A} + \beta_{i}^{AB} + \beta_{j}^{AD} + \beta_{k}^{AC} + \beta_{jk}^{ACD},$$

gdzie efekty poszczególnych czynników sumują się do zera (np. $\beta_{1}^{AD} + \beta_{2}^{AD} = 0$); wynika to z ograniczeń omówionych w podpunkcie „Modele nasycone”.
Ważną cechą modeli logitowych, której nie można wywnioskować z równania, jest to, że trójczynnikowa interakcja wszystkich zmiennych niezależnych (B, C, D) występuje w modelu, tak samo jak wszystkie pochodne od niej efekty niższego rzędu, tzn. {BC}, {BD}, {CD}, {B}, {C}, {D}. Parametry wpływu tych czynników nie pojawiają się w równaniu logitowym szans oczekiwanych kategorii czynnika A (upraszczają się), ale wszystkie odpowiadające im rozkłady brzegowe muszą być dopasowane w procesie dobierania frekwencji oczekiwanych, za pomocą których określa się szanse. W każdym modelu logitowym musi być zachowana zgodność zaobserwowanych i oczekiwanych rozkładów brzegowych interakcji wszystkich zmiennych niezależnych, nawet jeśli nie jest ona statystycznie istotna. Ten fakt stanowi główną różnicę między uogólnionymi modelami log-liniowymi a logitowymi.

Estymacja parametrów równania logitowego rozpoczyna się, tak jak w uogólnionych modelach log-liniowych, dopasowaniem podtabel rozkładów brzegowych. Aby uzyskać wartości poszczególnych β, przekształcamy odpowiadające im τ zgodnie z zależnością: $\beta^{AX} = 2 \ln \tau^{AX}$ (X oznacza zmienną B, C lub D, mającą wpływ na A). Można do tego użyć programu statystycznego.
Parametry modelu logitowego można interpretować podobnie jak addytywne współczynniki w regresji liniowej. Wartości dodatnie β znaczą, że obecność zmiennej niezależnej (lub interakcji zmiennych niezależnych) powoduje wzrost

szans „wyższej" kategorii zmiennej zależnej, zaś wartości ujemne, że powoduje jej spadek.

Aby oszacować wpływ zmiennej niezależnej o wielu kategoriach, trzeba wziąć pod uwagę wszystkie β. Efekty interakcji można interpretować na więcej niż jeden sposób. Wyobraźmy sobie, że w równaniu logitowym określającym szanse pierwszej kategorii zmiennej A (niskiego poziomu stresu w pracy) mamy:

$\beta_{11}^{\mathrm{ACD}} = -0{,}036$ (a zatem $\beta_{12} = 0{,}036$), $\beta_{21}^{\mathrm{ACD}} = -0{,}288$ (a więc $\beta_{22}^{\mathrm{ACD}} = 0{,}288$)

oraz

$\beta_{31}^{\mathrm{ACD}} = -0{,}324$ (stąd $\beta_{32}^{ACD} = 0{,}324$).

Wówczas fakt, iż $\beta_{32}^{ACD} = 0{,}324$, może oznaczać, że trzecia kategoria czynnika C (wysoki poziom wsparcia społecznego) „podwyższa" poziom A (poziom stresu w pracy) bardziej dla drugiej niż dla pierwszej kategorii czynnika D (bardziej dla braku poczucia samoskuteczności niż dla istnienia poczucia samoskuteczności) lub, równie dobrze, że druga kategoria czynnika D (brak poczucia samoskuteczności) lepiej koreluje (dodatnio) z A (stresem w pracy) w trzeciej kategorii czynnika C (dla wysokiego poziomu wsparcia społecznego) niż w pozostałych jego kategoriach (przy mniejszym wsparciu społecznym). Współczynnik ten nie oznacza jednak, że wszystkie przypadki należące do trzeciej kategorii czynnika C i drugiej kategorii czynnika D (duże wsparcie społeczne i brak poczucia samoskuteczności) mają odpowiednio wyższy poziom czynnika A (wysoki poziom stresu w pracy) niż przypadki należące do kategorii czynników C i D odpowiednio 3 i 1 (duże wsparcie społeczne i poczucie samoskuteczności) ani też 2 i 2 (średni poziom wsparcia społecznego i brak poczucia samoskuteczności). Znaczy on tylko, że logarytm naturalny szans dla komórki o indeksie 2, 3, 2 (wysoki poziom stresu, duże wsparcie i brak poczucia samoskuteczności) jest większy, niż można by oczekiwać na podstawie modelu, który wyklucza ten efekt, to znaczy takiego, że

$\Phi_{231}^{A} = \beta^{A} + \beta_{2}^{AB} + \beta_{1}^{AD} + \beta_{3}^{AC}$ (gdzie $\beta^{A} = \beta_{1}^{A} = 1/\beta_{2}^{A}$, $\beta_{2}^{AB} = \beta_{12}^{AB} = 1/\beta_{22}^{A}$ itd.).

Posługując się modelami log-liniowymi należy pamiętać, że parametry wpływu ilustrują różnice warunkowych frekwencji, warunkowych szans i warunkowych proporcji szans. Na przykład posiadanie przez osobę pewnej cechy A może podwyższać szanse posiadania przez nią pewnej cechy B, jednakże szanse te obiektywnie mogą być bardzo niskie.

TESTOWANIE DOPASOWANIA

Jak wyznaczać modele pasujące do danych?

Dla dowolnej dwuwymiarowej tabeli kontyngencji można dobrać pięć różnych modeli opisujących związki zawartych w niej danych. Dla tablic wielu zmiennych tych modeli może być bardzo dużo. Ważnym pytaniem, na które musi odpowiedzieć badacz, jest to, który z modeli najlepiej opisuje dane. Aby rozwiązać ten problem, trzeba porównać wyestymowane frekwencje oczekiwane każdego z modeli z frekwencjami zaobserwowanymi. Służy do tego statystyka χ^2 Pearsona

$$(\chi^2 = \sum_{i,j} \frac{(f_{ij} - F_{ij})^2}{f_{ij}})$$

lub statystyka największej wiarygodności

$$L^2 : L^2 = 2 \sum f_{ij} \ln(f_{ij} / F_{ij}).$$

Obydwie statystyki mają asymptotyczny rozkład prawdopodobieństwa χ^2, z tą samą liczbą stopni swobody równą liczbie niezależnych parametrów wpływu τ spośród tych, które są równe jedności (czyli nie wywierają wpływu na frekwencje). W przypadku zmiennej o n kategoriach (nazwijmy ją A), gdzie $n \geq 2$, należy pamiętać, że $n - 1$ parametrów wpływu tej zmiennej jest niezależnych (bo $\tau^A = \tau_1^A = 1/(\tau_2^A \tau_3^A ... \tau_n^A)$), co daje $n - 1$ stopni swobody.
Statystyka L^2 jest preferowana w stosunku do χ^2, ponieważ:
– frekwencje oczekiwane są estymowane metodami największej wiarygodności,
– może być wyznaczona jednoznacznie dla testów niezależności w wielozmiennowych tabelach kontyngencji.
Im większa wartość statystyki L^2 w stosunku do liczby stopni swobody, tym bardziej frekwencje oczekiwane różnią się od zaobserwowanych, tak więc aby model mógł być zaakceptowany jako dobrze odzwierciedlający dane, stosunek wartości L^2 do liczby stopni swobody musi być dostatecznie mały.
Niebanalnym pytaniem jest, jak dobrać wielkość błędów I (α – *alfa*) i II (β – *beta*) rodzaju, aby z jednej strony nie wnioskować o relacjach, które nie mają wystarczająco dobrego potwierdzenia w danych, a z drugiej strony – aby nie ignorować gorzej potwierdzonych efektów, które jednak występują w populacji. W praktyce decyzję o zaakceptowaniu modelu zazwyczaj podejmuje się, jeśli prawdopodobieństwo popełnienia błędu I rodzaju leży w przedziale [0,1; 0,35].

Porównywanie modeli

Testowanie hipotez

Modele porównuje się parami, w celu zweryfikowania hipotez dotyczących efektów poszczególnych czynników na frekwencje komórkowe. Najbardziej podstawowe i najczęściej testowane hipotezy to hipoteza niezależności i hipoteza identyczności rozkładów brzegowych.

Hipoteza niezależności. Jeśli porównujemy pewien model z modelem różniącym się od niego jedynie brakiem efektu interakcji wybranych czynników i wartość statystyki L^2 różni się znacząco dla obu modeli, może to być spowodowane tylko tym, że różniący je parametr wpływu interakcji czynników odzwierciedla dużą ich kowariancję, co oznacza, że nie są one niezależne (to znaczy, że proporcje szans warunkowych kategorii tych zmiennych są różne od jedności). Modele porównujemy badając stosunek różnicy wartości statystyk L^2 do różnicy liczby stopni swobody. Jeśli wielkość tego stosunku jest znacząca na dobranym wcześniej poziomie istotności, odrzucamy hipotezę niezależności zmiennych.

Hipoteza identyczności rozkładów brzegowych. Porównujemy modele różniące się obecnością parametru wpływu wybranej zmiennej. Wartość wyrażenia jest funkcją szans kategorii tej zmiennej. Jeśli zaobserwowane szanse warunkowe różnią się od jedynki, oznacza to wywieranie wpływu przez zmienną. Istotność tego wpływu badamy analogicznie jak w przypadku hipotezy niezależności.

Przypadek dużych prób – odpowiednik współczynnika determinacji z regresji wielokrotnej. Problemem w doborze najlepszego modelu dla dużych prób jest fakt, iż wartość statystyki L^2 jest proporcjonalna do wielkości próby. Gdy mamy do czynienia z próbami o liczności setek tysięcy, okazuje się, że jedynym modelem, który możemy zaakceptować jako pasujący do danych, będzie model nasycony, nawet jeśli efekty pewnych interakcji wyższego rzędu będą bardzo małe. Aby uniknąć tego problemu, stosuje się odpowiednik współczynnika determinacji (R^2) dla regresji wielokrotnej. Wyodrębnia się tzw. model bazowy, z którym następnie porównuje się inne, bardziej kompleksowe modele i sprawdza, na ile lepiej niż on pasują one do danych. Porównywanie wartości statystyk L^2 modelu testowanego i bazowego wskazuje na znaczenie zmienności danych spowodowanej wpływem czynników nie zawartych w modelu bazowym. Jeśli udział wartości L^2 modelu bazowego wyjaśniony przez model alternatywny wynosi co najmniej 90%, orzeka się, że należy przyjąć model alternatywny. Do szacowania wyjaśnionego udziału używa się następującego wzoru (odpowiednika R^2):

$$\frac{(L^2 m.bazowego) - (L^2 m.alternatywnego)}{(L^2 m.bazowego)}$$

PRZYKŁAD KONKRETNEGO ZASTOSOWANIA

Potencjalna liczba zastosowań modeli log-liniowych jest właściwie nieskończona (tworzenie modeli przyczynowych, analiza zmian zachodzących w czasie itd.). Za ich pomocą można analizować każdą tabelę krzyżową. W tym rozdziale podejmuję pewne konkretne zagadnienie, a mianowicie wykorzystanie modeli log-liniowych (m.in. tzw. modelu przyczynowego) do analizy danych z badania Półtorak (2001b) dotyczącego związku wybranych czynników psychologicznych z intensywnością używania alkoholu przez studentów.
W omawianym badaniu wzięto pod uwagę następujące zmienne dychotomiczne: nadużywanie alkoholu (nie/tak), reaktywność emocjonalna (niska/ wysoka), lęk (niski/wysoki) oraz psychastenia (niska/wysoka)[4]. Zmienne oznaczono kolejno jako N, R, L i Pt (dane dotyczące frekwencji zaobserwowanych w poszczególnych grupach zawiera tabela 4). Weryfikowana hipoteza dotyczyła istnienia dodatniego związku[5] czynników psychologicznych (R, L, Pt) z nadużywaniem alkoholu (N)[6]. Dla tej hipotezy utworzono model przyczynowy, nie zakładając z góry niczego o wzajemnych powiązaniach czynników R, L i Pt, a jedynie bezpośredni wpływ każdego z nich na N. Ten model to {NR}{NL}{NPt}.

Tabela 4.
Tabela krzyżowa frekwencji zaobserwowanych z czynnikami: psychastenia, lęk, reaktywność emocjonalna i nadużywanie alkoholu

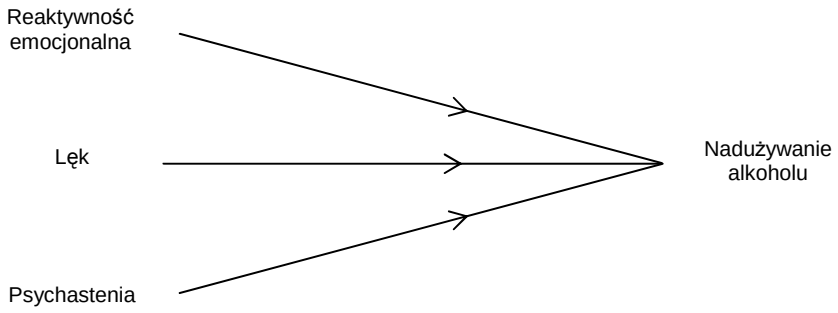
Psychastenia	Lęk	Reaktywność emocjonalna	Nadużywanie	
			nie	tak
Niska	niski	niska	10	35
		wysoka	17	8
	wysoki	niska	0	0
		wysoka	2	4
Wysoka	niski	niska	0	1
		wysoka	0	2
	wysoki	niska	0	0
		wysoka	5	7

[4] Szerzej zob. Półtorak, 2001b.

[5] Za „dodatni" związek uważany jest wtedy, kiedy odpowiadający mu współczynnik β jest dodatni.

[6] W związku tym, że w tabeli 4 występuje duża liczba komórek o zerowej frekwencji, do frekwencji we wszystkich komórkach tabeli dodano 0,01 (jest to w analizie log-liniowej powszechnie stosowana technika, gdy ma się do czynienia z komórkami o zerowych frekwencjach – zob. Goodman, 1970; Burke, Knoke, 1986); ze względu na minimalne różnice frekwencji nie zostało to zamieszczone w dodatkowej tabeli. Poziom istotności w analizie ustalono jako 0,05, zaś poziom p, powyżej którego program STATISTICA przyjmuje, że model pasuje do danych, jako 0,2.

Modele przyczynowe zazwyczaj reprezentowane są za pomocą równań lub grafów (zob. Burke, Knoke, 1986). Struktura takich grafów jest następująca: zmienne będące przyczynowymi poprzednikami umieszczone są zawsze po lewej stronie swoich następników. Kierunek strzałek jednostronnych prowadzi od przyczyny do skutku (jeśli dla pewnych zmiennych nie można powiedzieć, która jest przyczyną, a która skutkiem, łączy się je strzałkami dwustronnymi, co oznacza interakcje czynników; interakcje takie mogą się pojawiać tylko po lewej stronie diagramu). Diagram ilustrujący omawiany model przedstawia rysunek 1.



Rys. 1. Zakładany diagram przyczynowy

Analiza przyczynowa różni się od typowego modelowania logitowego, gdzie – tak jak w regresji liniowej – zazwyczaj mamy do czynienia z jedną zmienną zależną. Modelowanie przyczynowe składa się z serii niezależnych kroków, których rezultaty są zbierane i podsumowywane razem na końcu. Najpierw tworzona jest tabela kontyngencji dla zmiennych z lewej strony diagramu. Sprawdzane jest, czy są one skorelowane. Następnie szuka się najlepiej dopasowanego modelu dla tabeli poprzednio uwzględnionych zmiennych i ich pierwszego przyczynowego następnika. Kolejne kroki procedury są analogiczne (w ostatnim kroku dopasowuje się model do tabeli, w której uwzględnione są wszystkie zmienne). Ostatecznie otrzymywany model jest sumą modeli z obu etapów analizy.

Analiza przyczynowa dla omawianego modelu składała się z dwóch etapów:

– doboru najlepszego modelu dla trójwymiarowej tabeli kontyngencji złożonej ze zmiennych niezależnych R, L i Pt (przyczynowych poprzedników N);

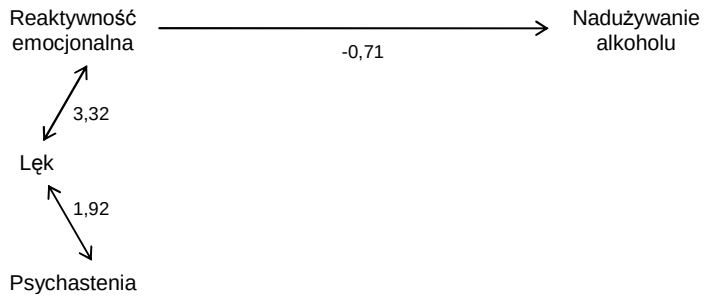
– doboru najlepszego modelu dla czterowymiarowej tabeli kontyngencji złożonej ze zmiennych niezależnych R, L i Pt oraz zmiennej zależnej N.

W pierwszym etapie najlepszy okazał się model {RL}{LPt}, co oznacza, iż istotne są interakcje między reaktywnością emocjonalną i lękiem oraz między lękiem i psychastenią. Końcowym modelem dla drugiego etapu jest {NR}{RLPt}, ostatecznie więc otrzymano model {NR}{RL}{LPt} z ośmioma stopniami swobody (8 = 2 + 6, gdzie 2 i 6 to liczby stopni swobody modeli od-

powiednio z pierwszego i drugiego etapu analizy) i wielkością statystyki L^2 równą 8,36 (8,36 = 1,13 + 7,23, gdzie 1,13 i 7,23 to wielkości statystyki L^2 dla modeli odpowiednio z pierwszego i drugiego etapu analizy). Tabela 5 zawiera dane dotyczące frekwencji oczekiwanych estymowanych przez ten model, zaś rysunek 2 przedstawia ilustrację graficzną otrzymanego modelu wraz z estymatorami parametrów wpływu[7] (siły związku) czynników.

Tabela 5.
Tabela krzyżowa frekwencji oczekiwanych na podstawie modelu {NR}{RL}{LPt} z czynnikami: psychastenia, lęk, reaktywność emocjonalna i nadużywanie alkoholu

Psychastenia	Lęk	Reaktywność emocjonalna	Nadużywanie	
			nie	tak
Niska	niski	niska	9,61	34,51
		wysoka	13,82	12,10
	wysoki	niska	0,003	0,01
		wysoka	3,21	2,81
Wysoka	niski	niska	0,42	1,50
		wysoka	0,60	0,52
	wysoki	niska	0,01	0,02
		wysoka	6,41	5,61



Rys. 2. Otrzymany diagram przyczynowy

Jak widać na rysunku, reaktywność emocjonalna wiąże się z nadużywaniem alkoholu, jednak jest to związek ujemny, a nie – jak zakładano – dodatni. Reaktywność emocjonalna jest dodatnio związana z lękiem, a lęk z psychastenią, dzięki czemu ostatecznie otrzymujemy, że nadużywanie alkoholu ujemnie wiąże się ze wszystkimi rozważanymi czynnikami psychologicznymi. Podsumowując można powiedzieć, iż powyższy model przyczynowy zupełnie zaprzecza weryfikowanej hipotezie.

[7] Są to współczynniki β z równania logitowego.

Dla lepszego zbadania tak zaskakującego rezultatu postanowiono wziąć pod uwagę również wpływ płci (wiadomo, że mężczyźni ogólnie częściej nadużywają alkoholu niż kobiety) i przyjrzeć się dokładniej kilku modelom dobrze opisującym dane zamieszczone w tabeli 6 (jest to tabela frekwencji zaobserwowanych z podziałem na płeć).

Tabela 6.

Tabela krzyżowa frekwencji zaobserwowanych z czynnikami: płeć, psychastenia, lęk, reaktywność emocjonalna i nadużywanie alkoholu

Płeć	Psychastenia	Lęk	Reaktywność emocjonalna	Nadużywanie	
				nie	tak
Kobieta	niska	niski	niska	5	5
			wysoka	13	4
		wysoki	niska	0	0
			wysoka	2	2
	wysoka	niski	niska	0	0
			wysoka	0	1
		wysoki	niska	0	0
			wysoka	2	3
Mężczyzna	niska	niski	niska	5	30
			wysoka	4	4
		wysoki	niska	0	0
			wysoka	0	2
	wysoka	niski	niska	0	1
			wysoka	0	1
		wysoki	niska	0	0
			wysoka	3	4

Wybrane modele uwzględniają wpływ płci oraz bardziej szczegółowe i bezpośrednie interakcje lęku i psychastenii z nadużywaniem alkoholu. Dane dotyczące liczby stopni swobody tych modeli, przybliżonych wartości statystyk L^2 i χ^2 oraz odpowiadających im poziomów p zamieszczono w tabeli 7.

Tabela 7.

Tabela wybranych modeli opisujących dane z tabeli 4 z liczbą stopni swobody (*df*), przybliżonymi wartościami statystyk L^2 i χ^2 oraz odpowiadającymi im poziomami *p*

Model	*df*	L^2	*p*	χ^2	*p*
{NPt}{NL}{NRP}{LRPt}	16	7,79	0,95	7,18	0,97
{NPtP}{NLRP}{LPt}	11	6,89	0,81	7,84	0,73
{NLP}{NRPtP}{LRPt}	9	4,75	0,86	4,07	0,91
{NRPt}{NRL}{NP}{LPt}{RP}	16	8,30	0,94	8,21	0,94
{PtP}{LP}{NRP}{LRPt}	16	8,36	0,94	7,27	0,97

Poniżej są zamieszczone niektóre wnioski, jakie można wysnuć na podstawie analizy tych modeli.

– nadużywanie alkoholu jest słabo dodatnio związane z psychastenią (β = 0,11);

– nadużywanie alkoholu właściwie nie wiąże się z lękiem (β = -0,04);

– nadużywanie alkoholu jest ujemnie związane z reaktywnością emocjonalną zarówno wśród mężczyzn, jak i wśród kobiet, aczkolwiek wśród mężczyzn silniej (odpowiednio β = -0,69 i -0,27);

– nadużywanie alkoholu jest dodatnio związane z psychastenią w przypadku kobiet (β = 0,65), zaś ujemnie w przypadku mężczyzn (β = -0,35);

– w przypadku wysokoreaktywnych emocjonalnie kobiet występuje dodatni związek między nadużywaniem alkoholu a lękiem (β = 0,59), wśród ogółu kobiet również, ale mniej silny (β = 0,41);

– w przypadku wysokoreaktywnych emocjonalnie mężczyzn występuje dodatni związek między nadużywaniem alkoholu a lękiem (β = 0,24), zaś ogólnie wśród mężczyzn – ujemny (β = -0,35);

– w przypadku wysokoreaktywnych emocjonalnie mężczyzn właściwie nie występuje związek między nadużywaniem alkoholu a psychastenią (β = 0,05);

– w grupie wysokoreaktywnych emocjonalnie kobiet występuje silniejszy dodatni związek między nadużywaniem alkoholu i psychastenią (β = 0,8) niż ogólnie w grupie kobiet (β = 0,65);

– wśród osób wysokoreaktywnych emocjonalnie występuje dodatni związek nadużywania alkoholu z lękiem (β = 0,49) i z psychastenią (β = 0,52);

– w podgrupie osób niskoreaktywnych emocjonalnie występuje ujemna korelacja między nadużywaniem alkoholu a poziomem lęku (β = -0,64);

– właściwie nie występuje korelacja między psychastenią a płcią (β = 0,02);

– występuje ujemna korelacja między lękiem a płcią (β = -0,24).

Podsumowując powyższe wnioski trzeba stwierdzić, iż wzięcie pod uwagę bezpośrednich interakcji nadużywania alkoholu z lękiem i psychastenią doprowadziło do stwierdzenia, iż między lękiem a nadużywaniem nie występuje właściwie żaden związek, zaś w przypadku psychastenii jest to związek bardzo słaby, ale dodatni. Oznacza to, że wysnuwanie wniosków o ujemnym związku tych czynników na podstawie pośredniego powiązania ich z nadużywaniem alkoholu poprzez reaktywność emocjonalną było zbyt pochopne. Ciekawy jest fakt potwierdzenia hipotezy o dodatnim związku nadużywania alkoholu z psychastenią i lękiem w podgrupie kobiet, a zaprzeczenia jej w podgrupie mężczyzn. Ciekawe jest również, iż wśród mężczyzn wysokoreaktywnych emocjonalnie niezbyt silny, ale dodatni związek między nadużywaniem alkoholu a lękiem występuje. Ogólnie w podgrupie osób wysokoreaktywnych emocjonalnie występują dodatnie związki między zmiennymi N i Pt oraz między zmiennymi N i L (warto zwrócić uwagę, że w grupie osób niskoreaktywnych związek między N i L również występuje i jest ujemny).

*

Chociaż artykuł ten nie jest szczegółowym opisem modelowania log-liniowego, daje pogląd na bogactwo i różnorodność możliwości stosowania tej metody w naukach społecznych. Wybór techniki analizy danych zawsze powinien być dostosowany do specyficznego sformułowania problemu badawczego, celem pracy nie było więc przekonanie czytelnika, iż ta metoda zawsze powinna być stosowana, gdy mamy do czynienia z tabelami kontyngencji, ale zapoznanie go ze stosunkowo nową i niezbyt jeszcze rozpowszechnioną metodą oraz pokazanie, że może być ona przydatna szczególnie wtedy, gdy analizowane zmienne są określone na skali nominalnej.

Ważne wydaje się zwrócenie uwagi czytelnika – zamierzającego w przyszłości korzystać z modelowania log-liniowego – na konieczność dobierania do badań dużych prób. Wynika to z faktu, iż w zależności od liczby analizowanych czynników i ich kategorii próba zostanie podzielona na bardzo wiele podgrup. Aby na podstawie rozkładów frekwencji zaobserwowanych w próbie rzetelnie szacować rozkłady w populacji, liczność wszystkich tych „najdrobniejszych" prób musi być odpowiednio duża[8].

Warto również zwrócić uwagę na to, iż często cenne może okazać się rozważenie większej liczby modeli niż tylko jednego (zwłaszcza iż zazwyczaj modeli dobrze pasujących do danych jest bardzo wiele). Pozwala to na głębsze i dokładniejsze zbadanie problemu oraz uniknięcie wysnuwania pochopnych wniosków.

BIBLIOGRAFIA

Burke, P. J., Knoke, D. (1986). *Log-linear models*. Sage university paper series on quantitative applications in social sciences. Beverly Hills–London: SAGE Publications.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, 65, 226-256.

Goodman, L. A. (1972). A general model for the analysis of surveys. *American Journal of Sociology*, 77, 1035-1086.

Półtorak, M. (2001a). *Zastosowania modeli log-liniowych w psychologii* (praca roczna na Wydziale Psychologii UW).

Półtorak, M. (2001b). *Związek wybranych czynników psychologicznych z intensywnością używania alkoholu przez studentów* (praca roczna na Wydziale Psychologii UW).

SPSS 8.0. Podręcznik użytkownika dołączony do programu.

StatSoft, Inc. (1995). *STATISTICA for Windows [Computer program manual]*.

[8] W opisanym przykładzie zastosowana mała liczebność próby uniemożliwiła wzięcie pod uwagę w analizie większej liczby czynników.