Psychologia Społeczna
2010 tom 5 2–3 (14) 199–216
ISSN 1896-1800

Zastosowanie regresji logistycznej w badaniach eksperymentalnych

Barnaba Danieluk
Instytut Psychologii UMCS, Lublin

W praktyce badawczej często spotykamy się z sytuacją, gdy mierzona przez nas zmienna zależna ma charakter zero-jedynkowy, przyjmując wartości 0 – brak czegoś i 1 – występowanie czegoś (konkretnego zachowania, zgody na coś, ujawnienia postawy, opinii itd.). Zarówno ogólny model liniowy, jak i analiza regresji liniowej nie znajdują zastosowania w sytuacji dychotomicznej, nominalnej zmiennej zależnej. W takiej sytuacji jesteśmy zmuszeni do stosowania analiz nieliniowych. Modelem regresyjnym stosowanym dla tego typu zmiennych zależnych jest regresja logistyczna. Artykuł prezentuje zastosowanie modelu dwumianowej regresji logistycznej w badaniach eksperymentalnych. Wyjaśnia specyfikę i sposób interpretacji charakterystycznych dla regresji logistycznej współczynników: ilorazów szans (*odds ratio*), współczynników Walda, ilorazów wiarygodności (*likelihood ratio*). Przybliża procedurę estymacji parametrów modelu metodą największej wiarygodności (*maximum likelihood*) oraz test dobroci dopasowania modelu Hosmera i Lemeshowa. W artykule zostały zawarte przykładowe analizy jednoczynnikowe (z predyktorem nominalnym i ilościowym), analiza dwuczynnikowa oraz analiza dwuczynnikowa z efektem interakcyjnym. Ograniczono do niezbędnego minimum liczbę wzorów i przekształceń algebraicznych, a same przykładowe analizy i ich interpretacje przeprowadzono krok po kroku z użyciem pakietu statystycznego SPSS w wersji 17.0 PL.

Słowa kluczowe: regresja logistyczna, dwumianowa regresja logistyczna, iloraz szans, iloraz wiarygodności, metoda największej wiarygodności, współczynnik Walda, SPSS

Planując badania eksperymentalne, psycholog-naukowiec wkracza na trudny i wymagający obszar metodologicznej poprawności. Zwykle najwięcej starań badacze wkładają w dobór zmiennych, ich operacjonalizację, zaplanowanie schematu eksperymentalnego, prawidłowy dobór próby i późniejszy losowy przydział do grup oraz kontrolę nad przebiegiem eksperymentu. Często zdarza się tak, że decyzję o wyborze metody obliczeniowej podejmują już po badaniu. Może się wtedy okazać, że dane uzyskane z badania nie pozwalają na zastosowanie najbardziej znanych i najczęściej stosowanych metod statystycznych, co czasami prowadzi do „naginania" danych.

Dla tego artykułu najistotniejsza będzie specyfika zmiennej zależnej mierzonej w eksperymencie. W psychologii społecznej szczególnie często mamy do czynienia z sytuacją, w której rezultatem oddziaływania eksperymentalnego jest wystąpienie (bądź brak wystąpienia) jakiegoś konkretnego zachowania. W najbardziej znanych eksperymentach dotyczących konformizmu, poznania społecznego, zmiany postaw, procesów grupowych, agresji, zachowań pro- i antyspołecznych, badacze poprzez manipulację warunkami eksperymentalnymi powodowali, że osoby badane godziły się na coś lub nie, przejawiały konkretne zachowanie lub się od niego powstrzymywały, ujawniały jakąś informację lub ją zatajały (por. Aronson, Wilson, Akert, 1997). Krótko mówiąc, mierzone zachowanie miało charakter zero-jedynkowy – albo uczestnik eksperymentu coś zrobił, albo tego nie zrobił. Oznacza to, że zmienna zależna w tych eksperymentach ma charakter nominalny, a konkretnie dychotomiczny. Zdarzają się oczywiście eksperymenty, w których badani pod wpływem czynników eksperymentalnych przejawiają jakościowo różne rodzaje zachowań (np. agresywne, uległe lub asertywne), lecz oznacza to tylko, że zmienna zależna ma w tym przypadku charakter politomiczny (czyli w dal-

Barnaba Danieluk, Instytut Psychologii UMCS, pl. Litewski 5, 20–080 Lublin, e–mail: barnaba@umcs.pl

szym ciągu jakościowy). Czasami eksperymentatorzy tak planują eksperymenty, aby uzyskać ilościową zmienną zależną. Na przykład w badaniach nad wpływem społecznym uległość bywa mierzona wielkością ofiarowanego datku wyrażoną w pieniądzach (por. Doliński, 2000). Jednak nie zawsze taki zabieg jest możliwy, oprócz tego wydaje się, że w wielu przypadkach dla weryfikacji hipotezy ważniejszy jest fakt, czy osoba badana zdecydowała się na dane zachowanie niż to, jak bardzo się w nie zaangażowała.

Stosowanie analizy wariancji w sytuacji, gdy zmienna zależna ma charakter nominalny, jest błędem z kilku powodów. Po pierwsze, warunkiem stosowania analizy wariancji jest minimum interwałowy poziom pomiaru tej zmiennej. Traktowanie zmiennych dychotomicznych jako mierzonych na skali interwałowej (przyjmujących wartości 0 i 1) jest poważnym błędem zniekształcającym rzeczywiste relacje między zmiennymi. W takiej sytuacji następuje przeszacowanie siły związku między zmiennymi, co prowadzi do błędu pierwszego rodzaju (Ferguson, Takane, 1999). Co więcej, stosowanie tego modelu wymaga równego rozkładu wariancji w grupach, czyli tzw. homogeniczności wariancji. Nie zawsze ten warunek jest spełniony, a przy zmiennych zależnych nominalnych jest niemożliwy do osiągnięcia (Stanisz, 2000). Najbardziej problematyczne jest jednak traktowanie zmiennej dychotomicznej jako zmiennej ciągłej. Zmienna zero-jedynkowa (np. uległość) może przyjąć dwie i tylko dwie wartości, bo przecież nie można się „trochę zgodzić" a „trochę nie zgodzić" na prośbę eksperymentatora, więc wszystkie pośrednie wartości dla tej zmiennej (np. 0,4) są nieosiągalne w rzeczywistości.

Problem stosowania metody ANOVA dla dychotomicznych zmiennych zależnych podjął na początku lat 70. XX w. Lunney (1970). Dowodził on, że po zakodowaniu zmiennej dychotomicznej zero-jedynkowo uzyskiwane w poszczególnych grupach średnie można traktować jako prawdopodobieństwo uzyskania przez zmienną zależną wartości 1. Zdawał sobie sprawę, że dla dychotomicznej zmiennej wariancja jest bezpośrednią funkcją średniej arytmetycznej, a przez to różne prawdopodobieństwa osiągnięcia przez zmienną zależną wartości 1 w poszczególnych porównywanych grupach jest równoznaczne z brakiem homogeniczności wariancji. Będąc świadomy łamania założeń analizy wariancji, przeprowadził procedurę Monte Carlo, aby porównać rozkład współczynników F z tysiąca symulowanych rozkładów z rozkładem teoretycznym F. Schematy, które wprowadził do procedury Monte Carlo, obejmowały analizy jedno-, dwu- i trójczynnikowe, w których za każdym razem zmienna zależna była dychotomiczna i zakodowana jako 0 i 1. Uzyskana zgodność rozkładów F w zakresie najwyższych percentyli pozwoliła mu na wyciągnięcie wniosków, że stosując analizę wariancji dla dychotomicznych zmiennych wynikowych, nie popełniamy błędu pierwszego rodzaju, o ile zadbamy o spełnienie kilku warunków. Pierwszym i najważniejszym jest równoliczność porównywanych grup (procedura Lunneya obejmowała wyłącznie schematy skorygowane). Drugim jest zachowanie minimum 20 stopni swobody dla wariancji błędu (jeżeli w grupie, w której wystąpiło najmniej zachowań kryterialnych – jedynek, proporcja tych zachowań wynosi więcej niż 0,2). Jeżeli natomiast proporcja w grupie, w której zaobserwowano najmniej zachowań kryterialnych jest bardziej skrajna (tj. mniejsza od 0,2), minimalna liczba stopni swobody dla wariancji wewnątrzgrupowej musi wynieść 40.

Cytowany powyżej artykuł Lunneya stał się dla wielu badaczy podstawą do stosowania analizy wariancji dla dychotomicznych zmiennych zależnych. Opierając się na nim, spełniając niezbyt wyśrubowane założenia, można było w prosty sposób stosować popularną analizę statystyczną do mniej „standardowych" danych. Jednak już rok po ukazaniu się artykułu Lunneya na łamach tego samego czasopisma (*Journal of Educational Measurement*) opublikowany został krytyczny artykuł D'Agostino (1971). Dowodzi on, że wprowadzone przez Lunneya do procedury Monte Carlo dane były symetryczne w obrębie rzędów tabeli krzyżowej (podobne proporcje w poszczególnych celkach), co wyrównało wariancję, a przez to spowodowało, że rozkład statystyki F pozostał niezaburzony. Udowadniając Lunneyowi tendencyjność, zaproponował bardziej poprawną procedurę stosowania ANOVY dla dychotomicznych zmiennych zależnych. Opierając się na analizie statystycznej i przekształceniach algebraicznych testu chi^2, pokazał, że stosowanie surowych zero-jedynkowych zmiennych nie oznacza błędu jedynie dla analiz jednoczynnikowych, o ile zastosowano dużą próbę (jej liczebności jednak D'Agostino nie precyzuje). Dla modeli dwuczynnikowych bez efektu interakcyjnego dopuszcza stosowanie surowych danych zero-jedynkowych, jeżeli w żadnej z podgrup proporcja nie wykracza poza przedział 0,25–0,75. Jednakże za najbardziej poprawne autor ten uważa stosowanie danych dychotomicznych po odpowiednim przekształceniu matematycznym. Proponuje on przekształcenie *arcus sinus*, a jako najbardziej poprawne – przekształcenie logitowe, czyli oparte na wyrażeniach logarytmicznych.

Dlaczego regresja logistyczna?

Powyższa dyskusja straciła obecnie rację bytu. Toczyła się ona w czasach, gdy badacze nie dysponowali alterna-

tywą wobec analizy wariancji przy stosowaniu dychotomicznych zmiennych zależnych. Wszystko zmieniło się w latach 70. XX w., chociaż pierwsze prace na temat zastosowań funkcji logistycznej powstały już pod koniec XIX wieku w środowisku statystyków zajmujących się opisem właściwości demograficznych. Pełny model regresji logistycznej został opracowany dopiero w 1972 roku. Opisu tego dokonał D. J. Finney w pracy *Probit analysis* (za: Stanisz, 2000). Ta metoda statystyczna znajduje zastosowanie wszędzie tam, gdzie zmienna zależna mierzona jest na skali nominalnej i przyjmuje dwie wartości, kodowane jako 0 – brak wystąpienia pożądanego zjawiska i 1 – wystąpienie danego zjawiska (Hosmer i Lemeshow, 2000). Istnieje również zmodyfikowana wersja klasycznej regresji logistycznej stosowana przy wielokategorialnych zmiennych zależnych – nazywana wielomianową regresją logistyczną, wykracza jednak poza ramy niniejszego opracowania.

Najbardziej znana i najprostsza metoda testowania istotności różnic między grupami dla zmiennych kategorialnych – test chi² – znajduje zastosowanie przede wszystkim w tabelach czteropolowych. Test chi^2 można stosować również dla tabel wielopolowych, lecz jego wynik staje się w takiej sytuacji trudno interpretowalny (istotność testu dotyczy całej tabeli – czyli wszystkich wartości obu zmiennych). Co więcej, często zdarza się tak, że jedna lub dwie celki w tabeli wielopolowej decydują o istotności statystycznej chi^2, podczas gdy w pozostałych polach liczebności pozostają zbliżone. Nie spełnia przez to wymogów stawianych przez badaczy planujących eksperymenty bardziej złożone niż obejmujące jedną dwukategorialną zmienną niezależną. Niezbędna staje się metoda pozwalająca na całościową analizę modelu, a więc uwzględniająca jednocześnie kilka zmiennych niezależnych, niekoniecznie tego samego typu.

Wymogi te spełnia regresja logistyczna. Jest to model matematyczny, którego możemy użyć w celu opisania wpływu jednej lub kilku zmiennych niezależnych na dychotomiczną zmienną zależną. Pozwala na włączenie do modelu zmiennych niezależnych o charakterze ilościowym (mierzonych na skali interwałowej) oraz jakościowym (mierzonych na skali nominalnej). Warunki stosowania tej metody obliczeniowej są znacznie mniej restrykcyjne niż Ogólnego Modelu Liniowego. Oprócz wspomnianej wcześniej dychotomiczności zmiennej zależnej, warunkiem użycia regresji logistycznej jest dostatecznie duża liczebność próby. Liczebność (n) próby musi być większa niż $10 \cdot (k+1)$, gdzie k jest liczbą zmiennych niezależnych (Stanisz, 2000).

W dalszej części artykułu zostanie opisany klasyczny, dwumianowy model regresji logistycznej w postaci jednoczynnikowej, dwuczynnikowej oraz dwuczynnikowej z efektem interakcyjnym. Liczba zaprezentowanych wzorów oraz opis procedur obliczeniowych został zminimalizowany do wartości ułatwiającej zrozumienie istoty regresji logistycznej, specyficznej dla niej metody estymacji parametrów, procedury testowania modelu i jej charakterystyczne wskaźniki. Zainteresowani procedurami obliczeniowymi oraz przekształceniami wzorów znajdą je w pracy Hosmera i Lemeshowa (2000).

Celem autora było dostarczenie wiedzy niezbędnej do samodzielnego i świadomego korzystania z regresji logistycznej bazującej na oprogramowaniu statystycznym. Pakietem statystycznym użytym do analizy danych opisanych w niniejszym artykule był program SPSS w wersji 17.0 PL. Również czytelnicy dysponujący wcześniejszymi wersjami tego programu będą mogli skorzystać ze wskazówek zawartych w tekście, o ile pracują na jego wersjach 12.0 i wyższych. Dotyczy to również nowszych niż 17.0 wersji pakietu SPSS, który obecnie został przemianowany na PASW. Czytelnicy korzystający z pakietu STATISTICA informacje niezbędne do stosowania analizy regresji logistycznej w tym programie odnajdą w pracy Stanisza (2000). Ponieważ autor ten opisuje model logistyczny bez efektów interakcyjnych, niniejszy artykuł pozostaje przydatny również dla użytkowników pakietu STATISTICA. Bazy danych wykorzystane do przykładowych obliczeń zostały umieszczone na serwerze UMCS, a ich dokładne adresy url zostały podane w dalszej części tekstu.

Model regresji logistycznej

Cały rachunek regresyjny opiera się na równaniach funkcji matematycznych. Rodzaj regresji wynika z rodzaju funkcji, na której jest oparta. Najpowszechniej stosowana regresja liniowa wykorzystuje równanie prostej:

$$(1) \quad f(x) = ax + b$$

W modelu regresji liniowej do zbioru danych empirycznych (naniesionych w postaci punktów na układ współrzędnych) dopasowywana jest metodą najmniejszych kwadratów linia prosta najlepiej obrazująca zależność między zmiennymi. Linia prosta stanowi najwygodniejszy sposób przedstawienia relacji między zmiennymi w badaniach psychologicznych, nie zawsze jednak oddaje rzeczywiste relacje. W wielu przypadkach zależności między zjawiskami nie mają charakteru prostoliniowego, lecz krzywoliniowy (typowym przykładem jest prawo Yerkesa-Dodsona). Model regresji nieliniowej różni się od modelu liniowego rodzajem funkcji matematycznej (regresja nieliniowa obejmuje takie rodzaje funkcji mate-

matycznych, jak funkcja potęgowa, kwadratowa, wykładnicza, wielomianowa, hiperboliczna) oraz metodą estymacji parametrów funkcji (przy regresji logistycznej stosuje się metodę maksymalnej wiarygodności – ang. *maximum likelihood*). Najprościej rzecz ujmując, w modelu regresji nieliniowej do danych empirycznych dopasowywana jest krzywa (wyrażona równaniem funkcji matematycznej), która najlepiej obrazuje zależność między zmiennymi.

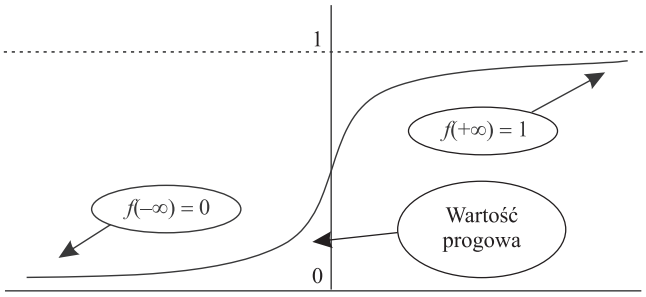
Ponieważ regresja logistyczna oparta jest na wyrażeniach logarytmicznych, warto w tym miejscu przypomnieć, czym jest logarytm i logarytm naturalny. Logarytm liczby N jest to wykładnik potęgi (x), do której należy podnieść stałą wartość podstawową (podstawę logarytmu – a), aby otrzymać N (Bronsztajn i Siemiendiajew, 1970). Wyrażenia logarytmiczne zapisuje się w postaci $\log_a N = x$, co jest równoważne z równaniem $a^x = N$. W rachunku prawdopodobieństwa i statystyce znajduje zastosowanie szczególny rodzaj logarytmu, a mianowicie logarytm naturalny, inaczej nazywany logarytmem Nepera lub hiperbolicznym. Jest to logarytm, w którym jako podstawę (a) stosuje się tzw. liczbę Eulera, symbolizowaną małą literą *e*. Wartość liczby Eulera wynosi w przybliżeniu 2,71828. Logarytm naturalny zapisuje się w postaci ln N, co jest równoważne z zapisem $\log_e N$, gdzie $e \approx 2{,}71828$. Jest to o tyle ważne, że analizując wydruki programów statystycznych dotyczące regresji logistycznej, spotkamy się często z wyrażeniem Exp (B), co oznacza funkcję wykładniczą wyrażenia B[1] o podstawie *e*, czyli e^B.

Model regresji logistycznej oparty jest na funkcji logistycznej. Funkcja ta określona jest wzorem (Hosmer i Lemeshow, 2000):

$$(2) \quad f(z) = \frac{e^z}{1+e^z}$$

Funkcja logistyczna przyjmuje wartości z przedziału < 0; 1 >, przy czym 0 i 1 są wartościami brzegowymi osiąganymi w +∞ i –∞ (Rysunek 1). Ma ona kształt wydłużonej litery S, a jej wartości dla wzrastających wartości x rosną od –∞ bardzo powoli, do momentu osiągnięcia wartości progowej. Po przekroczeniu wartości progowej wartości funkcji rosną gwałtownie, by ponownie ustabilizować się około wartości 1 (powoli zbliżając się do tej wartości w +∞).

Funkcja ta jest szczególnie przydatna przy analizie danych kategorialnych z dwóch powodów. Po pierwsze, przyjmuje wartości z przedziału < 0; 1 >, może więc opisywać wartości prawdopodobieństwa wystąpienia bądź niewystąpienia jakiegoś zjawiska (prawdopodobieństwo przyjmuje wartości z przedziału 0–1). Po drugie, zmienna zależna dychotomiczna przyjmuje tylko dwie wartości



Rysunek 1.
Postać funkcji logistycznej.

(kodowane najczęściej 0 i 1), przy czym pierwsza wartość oznacza zwykle brak występowania jakiegoś zjawiska (np. brak uległości na wpływ), a druga oznacza, że dane zjawisko miało miejsce (np. fakt uległości wobec wywieranego wpływu społecznego).

Równanie regresji logistycznej, podobnie jak równanie regresji liniowej (Ferguson i Takane, 1999) pozwala na obliczenie wartości oczekiwanej zmiennej zależnej. Ponieważ model regresji logistycznej dotyczy dwukategorialnych zmiennych zależnych (czyli przyjmujących jedynie dwie wartości: 0 i 1), wartość oczekiwana zmiennej zależnej Y' została zastąpiona wartością warunkowego prawdopodobieństwa, że zmienna zależna Y przyjmie wartość 1 dla zmiennych niezależnych $x_1, x_2, \ldots x_k$. Z własności funkcji logistycznej wynika, że obie te wartości (wartości oczekiwane zmiennej Y' oraz warunkowe prawdopodobieństwo przyjęcia wartości 1) są równe. Stąd model regresji logistycznej można wyrazić równaniem (Kleinbaum i Klein, 2002):

$$(3) \quad \mathrm{P}(\mathrm{Y}' = 1 \mid x_1, x_2, \ldots x_k) = \frac{e^{\alpha + \sum \beta_i x_i}}{1 + e^{\alpha + \sum \beta_i x_i}}$$

gdzie: $\mathrm{P}(\mathrm{Y}' = 1 \mid x_1, x_2, \ldots x_k)$ – warunkowe prawdopodobieństwo osiągnięcia przez zmienną zależną wartości 1 przy konkretnych wartościach zmiennych $x_1, x_2, \ldots x_k$
α – stała regresji dla regresji logistycznej
β_i – współczynnik regresji logistycznej dla i-tej zmiennej niezależnej
x_i – i-ta zmienna niezależna

Metody estymowania parametrów i testowania hipotez

Metoda największej wiarygodności (ML – *maximum likelihood*)

Dla obliczenia współczynnika P z równania (3) niezbędne jest oszacowanie wielkości stałej regresji dla regresji logistycznej (α) oraz współczynników regresji

logistycznej (β_i). W modelu regresji liniowej stała regresji (α) oraz współczynnik regresji (β) estymowany jest metodą najmniejszych kwadratów. Metoda ta nie znajduje zastosowania w przypadku regresji logistycznej ze względu na brak liniowości rozkładu zmiennej zależnej. Współczynniki regresji logistycznej estymowane są metodą największej wiarygodności (ang. *maximum likelihood*). Algorytm obliczeniowy metody największej wiarygodności opiera się na wielokrotnym estymowaniu wszystkich współczynników regresji, tak by zmaksymalizować prawdopodobieństwo uzyskania takich wyników, jakie osiągnięto w badanej próbie (Hosmer i Lemeshow, 2000). Wzór obliczeniowy uwzględnia łączone prawdopodobieństwo dla przypadków kryterialnych (czyli dla tych, dla których zmienna zależna osiągnęła w próbie wartość 1) i przypadków niekryterialnych (tych, dla których zmienna zależna wynosiła 0). Ta metoda estymacji parametrów wymaga bardzo żmudnych i skomplikowanych obliczeń (oznacza wielokrotne mnożenie współczynników prawdopodobieństwa dla różnych wartości parametrów aż do osiągnięcia największego iloczynu – maksymalnej wiarygodności), obecnie wykorzystuje się więc w tym celu odpowiednie oprogramowanie statystyczne.

Istnieją dwie odrębne formuły estymowania współczynnika największej wiarygodności (Kleinbaum i Klein, 2002). Metoda bezwarunkowa (*unconditional*) jest metodą prostszą i stosuje się ją w sytuacjach, gdy liczba zmiennych w modelu jest stosunkowo niewielka w porównaniu do liczebności próby. Metodę warunkową (*conditional*) stosujemy dla danych zależnych i wszędzie tam, gdzie liczba zmiennych w modelu jest stosunkowo duża. Literatura przedmiotu nie podaje konkretnej wartości, przy której zmiennych w modelu jest już zbyt dużo, aby można było stosować formułę bezwarunkową. Jednakże w typowych dla psychologii eksperymentalnej sytuacjach, czyli przy danych niezależnych, kilku predyktorach oraz efektach interakcyjnych możemy z powodzeniem stosować model bezwarunkowy. Kleinbaum proponuje (Kleinbaum i Klein, 2000), aby w sytuacjach dyskusyjnych stosować model warunkowy, gdyż jest modelem nieobciążonym. Jeszcze kilka lat temu pakiety statystyczne nie pozwalały na proste stosowanie formuły warunkowej (np. w pakiecie SPSS należało po specjalnym przekształceniu danych stosować regresję Coxa należącą do statystyk z grupy analiz przeżycia, gdyż standardowa regresja logistyczna w tym programie umożliwiała jedynie stosowanie modelu bezwarunkowego). Obecnie problem ten przestał istnieć – najnowsze wersje pakietu SPSS oferują dwie metody doboru zmiennych do modelu oparte na warunkowych ocenach parametrów ilorazu wiarygodności.

Iloraz wiarygodności (LR – *likelihood ratio*)

Wynikiem estymacji przeprowadzonej metodą największej wiarygodności jest: (a) wartość współczynnika największej wiarygodności (oznaczonego jako L), (b) macierz wariancji – kowariancji dla wyestymowanych współczynników regresji (przydatna przede wszystkim przy obliczaniu przedziałów ufności) oraz (c) lista zmiennych w modelu z odpowiadającymi im wyestymowanymi współczynnikami regresji oraz wartościami błędów standardowych (Kleinbaum i Klein, 2002).

W tym momencie najważniejsze dla badacza będzie to, czy zmienne wprowadzone do modelu istotnie wpływają na zmienną wynikową (zależną). Literatura opisuje dwa podejścia do testowania istotności współczynników regresji logistycznej: obliczanie ilorazu wiarygodności (*likelihood ratio* – LR) dla całego modelu oraz obliczanie wartości statystyki Walda dla każdej składowej osobno (zmiennej w modelu lub ich interakcji).

Pod względem statystycznym bardziej poprawne jest stosowanie obliczania ilorazu wiarygodności, gdyż uwzględnia on istotność całego modelu, a nie tylko pojedynczych, wyizolowanych parametrów. Stosowanie ilorazu wiarygodności nie wyklucza oczywiście obliczania statystyk Walda dla konkretnych zmiennych, gdyż dzięki tej ostatniej możemy ocenić, które z wprowadzonych zmiennych niezależnych w największym stopniu wpłynęły na zmienną zależną.

Stosując iloraz wiarygodności, odpowiadamy na pytanie, czy model zawierający zmienną (zmienne) niezależne da nam lepsze przewidywanie wyników (czyli np. zachowania badanego) niż model niezawierający tej (tych) zmiennej(ych). Obliczanie tego współczynnika oznacza za każdym razem porównanie dwóch wartości statystyki wiarygodności, a konkretnie jej szczególnej postaci, czyli zlogarytmizowanej wartości statystyki wiarygodności pomnożonej przez wartość –2 (–2 logarytm wiarygodności – *log likelihood*). Nie jest przedmiotem tego artykułu szczegółowe wyjaśnianie, dlaczego współczynnik wiarygodności przedstawiany jest w takiej formie, zainteresowanych odsyłam do pracy Hosmera i Lemeshowa (2000). Za wyjaśnienie niech posłuży fakt, że rozkład współczynnika wiarygodności w tej postaci pokrywa się z rozkładem chi^2 i dzięki temu jest dosyć łatwy w interpretacji. Warto w tym miejscu dodać, że logarytm wiarygodności jest odpowiednikiem sumy kwadratów dla reszt z regresji liniowej, to znaczy informuje o tym, jak wiele informacji o wariancji zmiennej zależnej pozostaje niewyjaśnionych po dopasowaniu modelu regresyjnego (Field, 2005). Stąd wysokie wartości logarytmu wiarygodności oznaczają słabo dopasowany model regresyjny, gdyż im wyższa jego wartość, tym więcej zmienności zmiennej zależnej

pozostaje niewyjaśnionych. Obliczając iloraz szans, porównujemy logarytm wiarygodności dla modelu zredukowanego (mniejszy model, zawierający mniejszą liczbę zmiennych niezależnych) z logarytmem wiarygodności dla modelu pełnego (większy model, zawierający więcej zmiennych niezależnych). Zwykle porównujemy modele różniące się od siebie jedną zmienną niezależną, po to, by sprawdzić, czy dodana zmienna istotnie zwiększa trafność przewidywań modelu (Kleinbaum i Klein, 2002). Wzór na iloraz wiarygodności ma postać:

$$(4) \quad LR = -2lnL_1 - (-2lnL_2)$$

Rozkład wartości ilorazu wiarygodności jest zgodny z rozkładem chi^2 z tyloma stopniami swobody, iloma zmiennymi różniły się model pełny od modelu zredukowanego. Do niedawna obliczanie ilorazu wiarygodności w pakietach statystycznych polegało na ręcznym definiowaniu modelu pełnego i zredukowanego, wprowadzaniu obu metodą krokową, a następnie odczytywaniu wartości logarytmu wiarygodności, odejmowanie tych wartości, aby w końcu otrzymaną różnicę odnieść do tablic rozkładu chi^2 (pamiętając, że różnica –2 logarytmu wiarygodności ma rozkład chi^2 z liczbą stopni swobody równą liczbie parametrów różniących model pełny od modelu zredukowanego). Obecnie mamy do dyspozycji metody selekcji postępującej i eliminacji wstecznej oparte na ilorazie wiarygodności (cała procedura została zautomatyzowana).

Współczynnik Walda (Z-Walda)

Inną metodą testowania hipotez w regresji logistycznej jest współczynnik Walda (Z). Stosowany jest do testowania hipotez zerowych dla współczynników regresji logistycznej każdej zmiennej w modelu (hipotez o zerowej wartości współczynnika regresji, czyli o braku wpływu predyktora na zmienną wynikową H_0: $\beta_i = 0$). Rozkład współczynnika Walda jest w przybliżeniu zgodny z rozkładem normalnym w dużych próbach. Natomiast rozkład współczynnika Walda podniesionego do kwadratu (Z^2) zgodny jest z rozkładem chi^2 z jednym stopniem swobody. W większości programów statystycznych (również w SPSS) współczynnik Walda przedstawiony jest w formie podniesionej do kwadratu, stąd często mówi się o współczynniku chi^2 Walda. Wzór obliczeniowy dla współczynnika Walda jest bardzo prosty i opiera się na już wyestymowanych współczynnikach regresji oraz ich błędach standardowych (Hosmer i Lemeshow, 2000):

$$(5) \quad Wald(Z) = \frac{\beta_i}{SE_\beta}$$

Jednoczynnikowa analiza regresji logistycznej dla predyktora dychotomicznego

Rozważmy prosty przykład jednoczynnikowej regresji logistycznej. Pod adresem http://spoleczna.umcs.lublin.pl/pliki/logistyczna1.sav znajduje się baza danych programu SPSS zawierająca dane z przykładowego eksperymentu z dziedziny wpływu społecznego. Wyobraźmy sobie prosty eksperyment, w którym próbowaliśmy przekonać badanych do spełnienia prośby albo formułując ją bez żadnego kontekstu (warunek kontrolny), albo poprzedzając ją dużą prośbą, na którą w zdecydowanej większości badani nie będą skłonni się zgodzić (warunek techniki „Drzwi zatrzaśniętych przed nosem"). W bazie danych znajdziemy zmienną niezależną DITF (od *Door In the Face*) przyjmującą dwie wartości: 0 dla grupy kontrolnej i 1 dla grupy eksperymentalnej oraz dychotomiczną zmienną zależną Uległość (o wartościach: 0 – brak zgody, 1 – zgoda).

System kodowania zmiennych nie jest sprawą dowolną. Zmienna zależna niekoniecznie musi zostać zakodowana jako 0 i 1, gdyż SPSS przekodowuje zmienną zależną zakodowaną w inny niż zero-jedynkowy sposób. Ważne jest, aby kategoria diagnostyczna miała wartość wyższą niż niediagnostyczna. W naszym przykładzie zatem moglibyśmy zakodować brak zgody jako 15, a zgodę jako 49 (gdyż SPSS przypisałby jedynkę wyższej wartości – czyli zgodzie), natomiast niepoprawne byłoby zakodowanie braku zgody jako 1, a zgody jako 2. Zmienna niezależna w naszym przykładzie również jest zmienną nominalną (i do tego dychotomiczną), dlatego także musi zostać zakodowana. Ponownie stosujemy kodowanie zero-jedynkowe (jest to o tyle ważne, że zakodowanie w inny sposób zmienia współczynniki regresji, a także wartość ilorazów szans dla zmiennych).

Z menu programu SPSS wybieramy Analiza → Regresja → Logistyczna, zmienną Uległość wprowadzamy jako zależną, zmienną DITF (Manipulacja) jako współzmienną. Dla jednoczynnikowej regresji logistycznej nie ma większego znaczenia, jaki sposób wprowadzania danych do modelu wybierzemy. Ponieważ jednak chcielibyśmy sprawdzić istotność współczynników regresji obiema metodami (poprzez iloraz wiarygodności i współczynnik Walda) wybieramy metodę selekcji postępującej opartej na ilorazie wiarygodności. Następnie uruchamiamy analizę. W edytorze raportów znajdujemy na początku informację o wprowadzonych danych oraz sposób, w jaki program zakodował wartości zmiennej zależnej (zwróćmy na to uwagę, pamiętając o informacjach z poprzedniego akapitu). Następnie SPSS generuje wyniki dla bloku (modelu) zerowego, czyli modelu zawierającego tylko i wy-

łącznie stałą regresji, z wyłączeniem wszystkich predyktorów. Dla modelu zerowego SPSS nie oblicza wartości statystyki wiarygodności, stąd nie znalazła się tam tabela „Model – podsumowanie".

Interesujące nas informacje znalazły się w kolejnym bloku (Bloku 1), opisującym model zawierający naszą zmienną niezależną. Tym razem mamy możliwość zapoznania się z wartością statystyki –2 logarytm wiarygodności, będącą odpowiednikiem statystyki R^2 w klasycznej regresji liniowej. Ponieważ wartość logarytmu wiarygodności nie jest tak intuicyjnie interpretowalna, jak wartość statystyki R^2, SPSS podaje wartości tzw. pseudo-R^2. Statystyka R^2 Coxa i Snella oparta jest na wartości logarytmu wiarygodności dla uzyskanego modelu, porównanego z logarytmem wiarygodności dla modelu zerowego z uwzględnieniem wielkości próby (Field, 2005). Ponieważ nie osiąga ona nigdy teoretycznego maksimum równego 1, program zawiera również modyfikację tego współczynnika w postaci R^2 Nagelkerkego. Sposób interpretacji tych współczynników jest analogiczny jak przy regresji liniowej, informując nas o stopniu, w jakim otrzymany model wyjaśnia wariancję zmiennej zależnej.

„Tabela klasyfikacji" pozwala wnioskować o stopniu dopasowania modelu do rzeczywistych danych. Są w niej zestawione wartości obserwowane z przewidywanymi na podstawie otrzymanego modelu. Dla naszych danych obserwujemy podobną trafność modelu w przypadku przewidywania zgody na prośbę (wartość kryterialna) i odmowy spełnienia prośby. Procent poprawnych klasyfikacji dla modelu wynosi w tym przypadku 68,3%.

Najważniejszą częścią wyników jest tabela „Zmienne w modelu", której budowa przypomina odpowiednią tabelę dla regresji liniowej. Zawarte są w niej współczynniki regresji, błędy standardowe tych współczynników oraz współczynniki Walda wraz z istotnościami.

Wnikliwy czytelnik (dysponujący kalkulatorem) może się zorientować, że wartość podawana przez SPSS jako współczynnik Walda jest tak naprawdę wartością chi^2 Walda, czyli współczynnikiem Walda podniesionym do kwadratu (Z^2). Z punktu widzenia testowania hipotezy o wpływie zmiennej niezależnej (manipulacji eksperymentalnej) na uległość odrzucamy hipotezę zerową o braku wpływu – która jest jednoznaczna z hipotezą o zerowej wartości współczynnika regresji dla tej zmiennej, i przyjmujemy hipotezę alternatywną, świadczącą o tym, że manipulacja eksperymentalna wpłynęła na uległość.

Podstawiając wartości estymowanych współczynników regresji do wzoru funkcji regresji logistycznej (3), otrzymujemy:

$$P(Y' = 1 | x_1) = \frac{e^{\alpha + \beta x_1}}{1 + e^{\alpha + \beta x_1}}$$

$$P(Y' = 1 | x_1) = \frac{e^{-0,693 + 1,54 x_1}}{1 + e^{-0,693 + 1,54 x_1}}$$

Aby obliczyć przewidywane (na podstawie modelu) prawdopodobieństwo wystąpienia zdarzenia kryterialnego zmiennej zależnej, musimy podstawić do powyższego wzoru konkretną wartość zmiennej niezależnej. Chcąc obliczyć oczekiwane prawdopodobieństwo uległości w grupie eksperymentalnej, za x_1 podstawiamy wartość 1 (zgodnie z tym, jak zakodowaliśmy zmienne). Za pomocą kalkulatora naukowego (musi zawierać funkcje logarytmiczne) otrzymano przewidywane prawdopodobieństwo uległości w grupie eksperymentalnej równe w przybliżeniu 0,70, czyli około 70%. W grupie kontrolnej (za x_1 podstawiamy 0) prawdopodobieństwo to wynosi 0,33, czyli około 33%.

Tabela 1.
Wartość logarytmu wiarygodności oraz pseudo-R^2 dla modelu jednoczynnikowego z predyktorem dychotomicznym

Model – podsumowanie

Krok	-2 logarytm wiarygodności	R kwadrat Coxa i Snella	R kwadrat Nagelkerkego
1	74,843	0,129	0,172

Tabela 2.
Wartości współczynników regresji dla modelu jednoczynnikowego z predyktorem dychotomicznym

Zmienne w modelu

		B	Błąd standardowy	Wald	*df*	Istotność	Exp(B)
Krok 1[a]	DITF	1,540	0,556	7,686	1	0,006	4,667
	Stała	–0,693	0,387	3,203	1	0,074	0,500

[a] – zmienne wprowadzone w kroku 1: DITF

Ostatnia tabela raportu – „Model po usunięciu składników" zawiera wyniki ilorazu wiarygodności (LR), porównującego w tym przypadku model pełny, zawierający zmienną niezależną z modelem zredukowanym, zawierającym tylko stałą regresji. Wartość ilorazu wiarygodności znajdziemy w kolumnie „Zmiana w wartości –2 logarytm wiarygodności", a wynosi ona 8,268 przy jednym stopniu swobody (model pełny od modelu zredukowanego różnił się jedną zmienną) i jest istotna statystycznie na poziomie $p = 0{,}004$. Można zauważyć, że wartość LR jest zbliżona do wartości statystyki Walda. Nie jest to przypadek, gdyż oba współczynniki testowały istotność jednego i tego samego parametru i oba współczynniki opierają się na rozkładzie chi^2.

Opisując zawartość tabeli „Zmienne w modelu", nie wspomniano o wartościach ostatniej kolumny, oznaczonej jako Exp(B), a równoznacznej z funkcją wykładniczą odpowiedniego współczynnika regresji o podstawie *e*. Jest to wartość o tyle ważna, że jest ona równa wartości ilorazu szans dla tego predyktora.

Pojęcie ilorazu szans (*odds ratio*)

W modelu regresji logistycznej, podobnie jak w regresji liniowej, podstawowe znaczenie mają wyestymowane wartości współczynników regresji logistycznej (β_1, β_2, …, β_k), stałej regresji (α) oraz ich statystyczna istotność. Współczynniki regresji logistycznej nie stanowią jednak miary, która w obrazowy sposób przedstawia zależność między zmiennymi. W modelu regresji liniowej jako miary wpływu zmiennych niezależnych na zależną używany jest współczynnik determinacji R^2. Model regresji logistycznej nie pozwala na oszacowanie współczynnika determinacji, umożliwia jednak obliczenie innego parametru – tzw. ilorazu szans (*odds ratio*). Pod pojęciem szansy rozumie się stosunek prawdopodobieństwa, że dane zjawisko wystąpi (np. że dana osoba spełni prośbę eksperymentatora) do prawdopodobieństwa, że dane zjawisko nie wystąpi (np. że dana osoba odmówi prośbie eksperymentatora). Szansę wystąpienia danego zjawiska w przypadku A określa się wzorem (Stanisz, 2000):

$$(6) \quad S(A) = \frac{p(A)}{p\,(\text{nie}\,A)} = \frac{p(A)}{1-p(A)}$$

Gdy w 30-osobowej grupie osób badanych 6 osób spełni prośbę eksperymentatora, prawdopodobieństwo wystąpienia zjawiska dla tej grupy p(A) wynosi 6/30 = 0,2; stąd szansa spełnienia prośby eksperymentatora w tej grupie wynosi S(A) = 0,2/(1–0,2) = 0,25, czyli ¼. Możemy więc powiedzieć, że prawdopodobieństwo spełnienia prośby eksperymentatora równa się ¼ prawdopodobieństwa odmówienia tej prośbie, ewentualnie, że prawdopodobieństwo odmowy prośbie eksperymentatora jest 4 razy większe niż prawdopodobieństwo jej spełnienia. Iloraz szans odnosi się do sytuacji, gdy występowanie danego zjawiska badane jest w dwóch niezależnych grupach. Wyraża się on stosunkiem szansy wystąpienia tego zjawiska w grupie A, czyli S(A), do szansy wystąpienia tego zjawiska w grupie B, czyli S(B). Wzór na iloraz szans przyjmuje postać:

$$(7) \quad OR_{A\times B} = \frac{S(A)}{S(B)} = \frac{p(A)}{1-p(A)} \div \frac{p(B)}{1-p(B)}$$

gdzie: $OR_{A\times B}$ – iloraz szans (*odds ratio*) dla grup A i B

Przykładowo, jeżeli oprócz wspomnianej wyżej 30-osobowej grupy badanych, w której tylko 6 osób uległo prośbie eksperymentatora, w planie badawczym uwzględniono drugą, 30-osobową grupę osób (np. w której prośbę eksperymentatora poprzedzono odpowiednią manipulacją eksperymentalną) i w grupie tej prośbie badacza uległo 22 uczestników, szansa spełnienia prośby eksperymentatora wynosi dla tej grupy S(B) = 0,73 / 0,27 = 2,70. Wartości szans podstawia się do wzoru na iloraz szans w ten sposób, że w liczniku znajduje się wartość szansy tej grupy, w której zakładamy oddziaływanie eksperymentalne. W naszym przypadku jest to grupa B, stąd iloraz szans dla grup B i A wynosi OR_{BxA} = S(B) / S(A) = 2,70 / 0,25 = 10,8. Oznacza to, że szansa na uzyskanie zgody na prośbę eksperymentatora jest prawie 11 razy większa w grupie z wprowadzoną manipulacją eksperymentalną niż w grupie pozbawionej manipulacji eksperymentalnej. Gdy otrzymany iloraz szans przekracza wartość 1, oznacza to, że szansa wystąpienia danego zdarzenia jest większa w grupie pierwszej (z licznika) niż w grupie drugiej (z mianownika). Przy wartościach OR < 1 zależność jest odwrotna.

Wzór (7) pozwala jedynie na obliczenie ilorazu szans dla próby. Jeżeli chcemy otrzymać wartości ilorazu szans dla populacji, musimy oprzeć się na wyestymowanych współczynnikach regresji. Obliczanie ilorazów szans w modelu regresji logistycznej opiera się na tzw. logitowej postaci funkcji logistycznej (Kleinbaum i Klein, 2002). Przedstawienie przekształcenia logitowego wraz z jego związkiem ze współczynnikiem ilorazu szans z pewnością podniosłoby świadomość metodologiczną Czytelnika, jednakże nie jest niezbędne do prawidłowego stosowania regresji logistycznej i trafnej jej interpretacji. Zainteresowanych odsyłam do opracowania Kleinbauma, w literaturze polskojęzycznej przekształcenie to znajdziemy w książce Stanisza (2000). Głównym powodem stosowania przekształcenia funkcji logistycznej w formę logitową jest umożliwienie wyrażenia tej funkcji w formie równania przedstawiającego zmienne niezależne w rela-

cji liniowej, mimo że obiektywnie ich zależność jest nieliniowa (Field, 2005). Jak było już wspomniane wcześniej, iloraz szans porównuje szansę wystąpienia zjawiska (czyli przyjęcia przez zmienną zależną wartości kryterialnej – 1) w dwóch grupach. Przy jednej dychotomicznej zmiennej niezależnej istnieje tylko jedno takie porównanie (zwykle oznacza ono porównanie grupy eksperymentalnej z grupą kontrolną). Jeżeli natomiast w naszym modelu jest więcej zmiennych niezależnych i do tego przyjmujących więcej wartości, możliwych do obliczenia ilorazów szans staje się tyle, ile istnieje możliwych kombinacji tych zmiennych dla dwóch grup. Gdybyśmy np. oprócz wspomnianej zmiennej wprowadzili do modelu trójkategorialną zmienną „nastrój" (pozytywny, neutralny i negatywny) moglibyśmy obliczyć trzy różne ilorazy szans zależne od nastroju badanych osób. Po dodaniu kolejnej zmiennej liczba możliwych do obliczenia współczynników OR wzrasta w zależności od liczby wartości, które przyjmuje ten nowy czynnik. Liczba możliwych współczynników OR staje się szczególnie duża, gdy jako predyktor włączymy do modelu zmienną ilościową (np. wynik standaryzowanego narzędzia psychologicznego). Ogólny wzór na iloraz szans dla dwóch dowolnie zdefiniowanych grup ma postać (Kleinbaum i Klein, 2002):

$$(8) \quad OR_{A\times B} = e^{\sum \beta_i (x_{Ai} - x_{Bi})}$$

gdzie: x_{Ai} oznacza wartość i-tej zmiennej niezależnej w grupie A
x_{Bi} oznacza wartość i-tej zmiennej niezależnej w grupie B
β_i oznacza współczynnik regresji dla i-tej zmiennej niezależnej

Warto zwrócić uwagę, że we wzorze na iloraz szans nie występuje stała regresji (α). Wzór (8) na iloraz szans znajduje zastosowanie w modelach wieloczynnikowej regresji logistycznej, ale bez efektów interakcyjnych.

Powróćmy jeszcze na chwilę do jednoczynnikowej regresji logistycznej i naszego prostego eksperymentu z „Drzwiami zatrzaśniętymi przed nosem". Jednoczynnikowa regresja logistyczna z dychotomiczną zmienną niezależną jest przypadkiem szczególnym, w którym wzór na iloraz szans (dla jedynego zresztą możliwego porównania grup) ma postać:

$$(9) \quad OR_{A\times B} = e^{\beta_1}$$

gdzie: β_1 oznacza współczynnik regresji dla predyktora

A ponieważ wyrażenie e^{β_1} jest równoznaczne z wyrażeniem exp (β_1), które znajdziemy w ostatniej kolumnie Tabeli 2 („Zmienne w modelu"), wiemy już, ile wynosi iloraz szans porównujący uległość w grupie eksperymentalnej i grupie kontrolnej. Na podstawie uzyskanego wyniku (OR = 4,667) możemy stwierdzić, że stosując technikę „Drzwi zatrzaśniętych przed nosem", mamy prawie pięciokrotnie większą szansę na uzyskanie zgody badanych w porównaniu do sytuacji, w której stosujemy tylko prośbę zasadniczą.

Jednoczynnikowa analiza regresji logistycznej dla predyktora ilościowego

Zanim wprowadzimy do modelu regresji logistycznej kolejne zmienne niezależne dla uzyskania modelu wieloczynnikowego, rozważmy przykład zmiennej niezależnej mierzonej na skali interwałowej. Jak było wspomniane, regresja logistyczna (podobnie jak liniowa) umożliwia wprowadzanie do modelu zmiennych mierzonych na różnym poziomie, zarówno jakościowych (nominalnych), jak i ilościowych (interwałowych).

Przypuśćmy, że badając skuteczność techniki „Drzwi zatrzaśniętych przed nosem" bierzemy pod uwagę możliwość, że uległość wobec próśb formułowanych przez nieznaną osobę na uczelnianym korytarzu może być modyfikowana przez poziom samooceny osób badanych. Dlatego w procedurze eksperymentalnej przewidzieliśmy pomiar samooceny skalą SES Rosenberga. Nie wdając się w szczegóły proceduralne tego fikcyjnego przecież eksperymentu, bierzemy pod uwagę dwie zmienne niezależne – analizowaną już wcześniej manipulację eksperymentalną (zmienną dychotomiczną) oraz poziom samooceny mierzonej skalą SES Rosenberga (zmienna interwałowa). Spróbujmy na początek wprowadzić do modelu regresji logistycznej tylko jeden predyktor – tym razem samą zmienną samooceny (oznaczoną w bazie danych jako SES). Ponownie wybieramy metodę selekcji postępującej opartej na ilorazie wiarygodności.

Współczynnik regresji β dla zmiennej Poziom samooceny okazał się istotny statystycznie $Z^2_{(1)} = 9{,}436$; $p = 0{,}002$. Wartość ilorazu szans dla tej zmiennej (OR = 0,847) nie jest już niestety tak łatwo interpretowalna, jak było to w przypadku zmiennej zero-jedynkowej. W tym miejscu konieczne jest dokładniejsze wyjaśnienie, czym jest iloraz szans zapisywany przez program SPSS w kolumnie Exp(B). Jest to iloraz szans dla jednostkowej zmiany wartości zmiennej niezależnej, czyli mówi o tym, jak zmienia się szansa wystąpienia zjawiska kryterialnego, gdy wartość zmiennej niezależnej wzrasta o 1. Przy dychotomicznym predyktorze taka jednostkowa zmiana wyczerpuje całą jego zmienność i oznacza zmianę z 0 do 1 (czyli najczęściej z grupy kontrolnej do eksperymentalnej). Dla naszej ilościowej zmiennej – czyli samooceny mierzonej skalą

Tabela 3.
Wartości współczynników regresji logistycznej dla modelu jednoczynnikowego z predyktorem interwałowym

Zmienne w modelu

		B	Błąd standardowy	Wald	*df*	Istotność	Exp(B)
Krok 1[a]	SES	–0,166	0,054	9,436	1	0,002	0,847
	Stała	4,651	1,519	9,377	1	0,002	104,665

[a] – zmienne wprowadzone w kroku 1: SES

SES – obliczony przez SPSS iloraz szans oznacza, że gdy wynik w teście samooceny Rosenberga wzrasta o 1 szansa spełnienia prośby eksperymentatora spada o ok. 15% (gdyż OR = 0,847, a więc jest niższe od 1). W większości przypadków iloraz szans dla zmiany jednostkowej nie obrazuje w sposób komunikatywny wpływu ilościowej zmiennej niezależnej. Wzrost wartości takiej zmiennej o jedną jednostkę jest na tyle mało znaczący, że trudny do interpretacji, poza tym przy zmiennych ilościowych o dużym rozstępie otrzymywane ilorazy szans dla zmiany jednostkowej są zwykle bardzo zbliżone do wartości 1. Dlatego bardziej użyteczne okazuje się obliczanie ilorazu szans dla większej niż jednostkowa zmiany predyktora ilościowego. Hosmer i Lemeshow proponują obliczanie ilorazu szans dla dowolnej zmiany wartości predyktora, przez pomnożenie wielkości tej zmiany przez uzyskany współczynnik regresji i podniesienie wartości liczby Eulera do uzyskanego iloczynu (Hosmer i Lemeshow, 2000), co można wyrazić wzorem:

$$(10) \quad OR(x) = e^{x\beta} = \exp(x\beta)$$

gdzie: OR(x) oznacza iloraz szans dla x-owej zmiany ilościowej zmiennej niezależnej

Gdybyśmy chcieli sprawdzić, jaki wpływ na prawdopodobieństwa spełnienia prośby eksperymentatora ma 10-punktowy wzrost wyniku w teście samooceny Rosenberga, wystarczyłoby podnieść podstawę logarytmu naturalnego (czyli *e*) do potęgi 10 · –0,166, co dałoby iloraz szans na poziomie OR(10) = 0,19. Oznacza to, że przy wzroście poziomu samooceny o 10 punktów w skali SES, szansa na uległość wobec prośby eksperymentatora maleje pięciokrotnie. Alternatywnie możemy skorzystać ze wzoru (8) i obliczyć iloraz szans dla dwóch grup o wybranych przez nas wartościach zmiennej niezależnej.

Popatrzmy jeszcze przez chwilę na ostatnią tabel naszego wydruku SPSS, prezentującą wyniki ilorazu szans dla modelu zawierającego poziom samooceny jako predyktor uległości na prośbę.

Różnica w wartościach statystyki –2 logarytm wiarygodności dla modelu pełnego, zawierającego zmienną Poziom samooceny w porównaniu z modelem zredukowanym (zawierającym wyłącznie stałą regresji), okazała się istotna statystycznie i, co ciekawe, wyższa niż w przypadku poprzednio analizowanego modelu zawierającego zmienną manipulacji eksperymentalnej techniką „Drzwi zatrzaśniętych przed nosem". Na pytanie o to, który z tych czynników w większym stopniu determinuje uległość, odpowiemy, stosując dwuczynnikową regresję logistyczną zawierającą oba analizowane predyktory.

Zanim przejdziemy do obliczania analizy dwuczynnikowej zaznaczmy, że zaprezentowany sposób wprowadzania do regresji logistycznej predyktora ilościowego zakłada liniową zależność między tą zmienną a logitową postacią zmiennej logistycznej. Ponieważ celowo nie zawarto w tym artykule szczegółów przekształcenia logitowego, powyższe założenie najłatwiej będzie zobrazować przykładem. W powyżej przedstawionym podejściu zakładamy, że jednostkowa zmiana predyktora spowoduje taką samą zmianę w prawdopodobieństwie wystąpienia zdarzenia kryterialnego niezależnie od bezwzględnej wartości predyktora. Odnosząc to do opisywanego przykładu, przyjmujemy, że wzrost wyniku w teście samooceny Rosenberga o 1 spowoduje taki sam spadek prawdopodobieństwa spełnienia prośby eksperymentatora bez

Tabela 4.
Wartości ilorazu wiarygodności dla modelu jednoczynnikowego z predyktorem interwałowym

Model po usunięciu składników

Zmienna		Logarytm wiarygodności modelu	Zmiana w wartości –2 logarytm wiarygodności	*df*	Istotność zmiany
Krok 1	SES	–41,555	12,087	1	0,001

względu na to, czy będzie to zmiana z wyniku 12 na 13 czy 39 na 40. Istnieją sposoby estymacji parametrów uwzględniające również zależność nieliniową, lecz wykraczają one zdecydowanie poza zakres niniejszego opracowania.

Dwuczynnikowa analiza regresji logistycznej

Dotychczas rozpatrywaliśmy model regresji logistycznej zawierający tylko jedną zmienną niezależną: nominalną (dychotomiczną) albo ilościową (interwałową). Teraz zajmiemy się modelem dwuczynnikowym, który w prosty sposób może być rozbudowany o kolejne czynniki, tworząc model wielozmiennowy. Zasady wprowadzania i analizy predyktorów są takie same dla dwóch, jak i dla wielu zmiennych, stąd przedstawiony przykład powinien umożliwić samodzielną analizę większych planów badawczych.

Powróćmy do naszego przykładu badań nad wpływem społecznym. Procedura eksperymentalna zakładała manipulację kontekstem formułowania prośby (bez prośby wstępnej – grupa kontrolna *vs.* wygórowana prośba wstępna – czyli technika „Drzwi zatrzaśniętych przed nosem"). Jednocześnie kontrolowaliśmy samoocenę badanych osób jako potencjalną zmienną modyfikującą uległość i podatność na wpływ społeczny. Oba te czynniki analizowane w modelu jednozmiennowym okazały się mieć istotny wpływ na poziom uległości osób badanych. Poprawność metodologiczna – w przeciwieństwie do analiz cząstkowych – wymaga stosowania analiz wielozmiennowych dlatego tym razem wprowadzimy do modelu regresji logistycznej obie zmienne.

W programie SPSS ponownie wybieramy z menu Analiza → Regresja → Logistyczna i jako współzmienne wprowadzamy zmienne DITF i SES. Zmienną zależną pozostaje Uległość. Na tym etapie metoda wprowadzania zmiennych do modelu zaczyna mieć większe znaczenie niż przy analizie jednozmiennowej. Metodę domyślnie proponowaną przez program SPSS (wprowadzania) wybieramy wtedy, gdy interesują nas skorygowane wartości ilorazów szans dla wszystkich zmiennych w modelu, bez względu na to, czy wszystkie predyktory przyczyniają się w sposób istotny do wyjaśnienia zmienności zmiennej wynikowej. Nas interesuje model zawierający tylko te zmienne, które zwiększają trafność przewidywań wartości zmiennej zależnej w porównaniu do modelu zredukowanego (zerowego), zawierającego wyłącznie stałą regresji. Dlatego ponownie wybieramy metodę selekcji postępującej opartej na ilorazie wiarygodności. Włącza ona kolejną zmienną do modelu, pod warunkiem że ta nowa zmienna powoduje istotny wzrost ilorazu wiarygodności.

Blok początkowy ponownie zawiera jedynie stałą regresji. W pierwszym kroku wprowadzona została zmienna poziomu samooceny (gdyż była bardziej istotna statystycznie). Ponieważ po wprowadzeniu zmiennej DITF (manipulacji eksperymentalnej) wartość ilorazu wiarygodności nie przekroczyła progu istotności domyślnie określonego na poziomie 0,05, drugi predyktor nie został uwzględniony w modelu. Okazało się, że model zawierający oba predyktory nie pozwalał przewidywać wartości zmiennej zależnej istotnie lepiej niż model zawierający tylko zmienną SES. Obrazuje to ostatnia tabela wydruku SPSS.

Niestety, SPSS nie podaje wartości ilorazu wiarygodności dla modelu zawierającego obie zmienne niezależne. Możemy je jednak uzyskać, definiując model regresji logistycznej ręcznie w formie krokowej (*stepwise*). W pierwszym bloku wprowadzamy model zawierający tylko zmienną zależną i poziom samooceny, w kroku drugim dodajemy zmienną manipulacji eksperymentalnej i uruchamiamy analizę. W edytorze raportów odnajdujemy wartość statystyki –2 logarytm wiarygodności dla modelu zredukowanego (zawierającego tylko zmienną SES). Wynosi on $-2 \log L_1 = 71{,}024$. Wartość tej samej statystyki dla modelu pełnego (zawierającego oba predyktory) wynosi $-2 \log L_2 = 67{,}443$. Po podstawieniu do wzoru (4) wartość ilorazu wiarygodności wynosi $LR = 3{,}581$ i przy jednym stopniu swobody (gdyż różnica między modelem pełnym a zredukowanym wynosi jedną zmienną) odnosimy ją do tablic rozkładu chi^2 znajdujących się w większości podręczników do statystyki. Po odnalezieniu odpowiedniego rzędu i kolumny w tablicy rozkładu chi^2 (Ferguson i Takane, 1999) okazuje się, że otrzymana wartość ilorazu wiarygodności nie przekracza wartości krytycznej testu równej 3,84 (dla poziomu istotności 0,05 i jednego stopnia swobody).

Okazuje się więc, że uległość wobec próśb jest w większym stopniu determinowana poziomem samooceny osób badanych niż stosowaną wobec nich techniką „Drzwi zatrzaśniętych przed nosem". Przy uwzględnieniu zmiennej samooceny okazało się, że to, czy badany został poddany procedurze „Drzwi zatrzaśniętych przed nosem" nie mia-

Tabela 5.
Zmienne niewłączone do modelu dla dwuczynnikowej regresji logistycznej

Zmienne niewłączone do modelu

			Ocena	*df*	Istotność
Krok 1	Zmienne	DITF	3,694	1	0,055
	Statystyki ogólne		3,694	1	0,055

ło takiego wpływu na uległość, jaki sugerowały wyniki jednoczynnikowej regresji logistycznej uwzględniającej wyłącznie manipulację eksperymentalną.

Nie oznacza to oczywiście, że stosowanie techniki DITF nie wpłynęło w żaden sposób na uległość badanych osób. Zmienna manipulacji eksperymentalnej została wyłączona z modelu dwuczynnikowego, chociaż jej wpływ na zmienną zależną niewiele odbiegał od wartości progowej (istotność na poziomie $p = 0{,}055$). W takich sytuacjach warto przeprowadzić dwuczynnikową analizę regresji metodą wprowadzania, gdy obie zmienne są „ręcznie" wprowadzone do modelu. W ten sposób otrzymujemy tzw. skorygowane współczynniki regresji logistycznej i odpowiadające im skorygowane ilorazy szans (Hosmer i Lemeshow, 2000). Taka analiza jest równoznaczna z analizą kowariancji, w której obie zmienne niezależne stanowią dla siebie nawzajem kowarianty. Wyjaśnijmy to na naszym przykładzie (patrz: Tabela 6).

W jednoczynnikowej analizie regresji przeprowadzonej dla zmiennej DITF (manipulacja eksperymentalna) współczynnik regresji wynosił $\beta = 1{,}54$, a iloraz szans dla grupy eksperymentalnej i kontrolnej wynosił OR = 4,667. Gdy włączyliśmy do modelu poziom samooceny (SES), stał się on dla zmiennej DITF kowariantem, a w wyniku tego współczynnik regresji i iloraz szans spadły odpowiednio do wartości $\beta = 1{,}127$ i OR = 3,118. Oznacza to, że gdy uwzględniamy wyłącznie manipulację eksperymentalną, szansa na uległość jest ponad 4,5-krotnie wyższa w grupie eksperymentalnej niż kontrolnej, natomiast przy kontroli poziomu samooceny ta szansa jest już tylko 3-krotnie większa, a wynik jest obarczony większym błędem statystycznym. Efekt, jaki wywiera zmienna SES na relację między manipulacją eksperymentalną a uległością, nazywany jest przez epidemiologów *confounding effect* i świadczy o tym, że zmienna SES jest skorelowana zarówno z manipulacją eksperymentalną, jak i ze zmienną zależną (Hosmer i Lemeshow, 2000).

Podobną rolę dla zmiennej SES pełni zmienna DITF w tym modelu. W modelu zawierającym wyłącznie poziom samooceny badanych współczynnik regresji dla tej zmiennej wynosił $\beta = -0{,}166$ a jednostkowy iloraz szans OR = 0,847. W momencie włączenia zmiennej DITF do modelu współczynniki te spadły odpowiednio do wartości $\beta = -0{,}138$ i OR = 0,871. Czyli wpływ poziomu samooceny na uległość przy kontroli warunku eksperymentalnego okazał się nieco mniejszy, choć zmiana ta jest praktycznie nieistotna.

Część badaczy (np. Menard) sugeruje, aby stosować metodę eliminacji wstecznej zamiast selekcji postępującej przy wprowadzaniu zmiennych do modelu. Uzasadniają to efektem tłumienia (*suppressor effect*), jaki jeden z predyktorów może wywierać na inny predyktor (Menard, 1995). Efekt ten polega na tym, że dana zmienna niezależna różnicuje wartości zmiennej zależnej tylko wtedy, gdy inna zmienna niezależna utrzymywana jest na tym samym poziomie. Czyli gdy obie zmienne wprowadzone są do modelu w jednym kroku (zmienna tłumiona i zmienna tłumiąca), stają się one dla siebie kowariantami, a przez to model ujawnia wpływ zmiennej tłumionej na zmienną wynikową (gdyż poziom zmiennej tłumiącej jest kontrolowany). Gdy natomiast zmienną tłumioną wprowadzi się do modelu osobno (we wcześniejszym kroku niż zmienna tłumiąca), może ona nie wykazać istotnego wpływu na zmienną zależną i w procedurze selekcji postępującej może zostać wykluczona z analizy.

Test dobroci dopasowania Hosmera i Lemeshowa

Testowanie modelu regresji logistycznej – bez względu na to, czy oparte na ilorazie wiarygodności (LR), czy na wartościach statystyki Walda – odpowiada na pytanie, czy model zawierający predyktory powoduje trafniejsze przewidywanie wartości zmiennej zależnej niż model zawierający mniej predyktorów. Zwykle modelem odniesienia jest model zerowy, zawierający wyłącznie stałą regresji, czyli oparty na samym rozkładzie zmiennej zależnej w próbie (bez uwzględnienia predyktorów). Gdy uzyskamy satysfakcjonującą wartość ilorazu wiarygodności, wiemy tylko, że nasz model zawierający X zmiennych niezależnych (model pełny) lepiej pozwala przewidywać wartości zmiennej zależnej niż model zredukowany, za-

Tabela 6.
Współczynniki regresji dla analizy dwuczynnikowej przeprowadzonej metodą wprowadzania

Zmienne w modelu

		B	Błąd standardowy	Wald	*df*	Istotność	Exp(B)
Krok 1[a]	DITF	1,137	0,603	3,555	1	0,059	3,118
	SES	–0,138	0,055	6,372	1	0,012	0,871
	Stała	3,299	1,608	4,211	1	0,040	27,092

[a] – zmienne wprowadzone w kroku 1: DITF, SES

wierający pewien podzbiór zmiennych X. Nie oznacza to jednak, że otrzymany model jest dobrze dopasowany do danych obserwowanych. Pewną informację niesie zawartość Tabeli klasyfikacji, znajdującej się w raporcie analizy regresji przeprowadzonej z użyciem programu SPSS. Przy każdym kroku (modelu) SPSS podaje informację o tym, jaki procent obserwacji z próby został poprawnie zakwalifikowany na podstawie modelu. Ma to jednak tylko wartość pomocniczą, gdyż nie można jednoznacznie stwierdzić, jaki procent trafnie zakwalifikowanych obserwacji oznacza dobre dopasowanie modelu.

Hosmer i Lemeshow zaproponowali procedurę szacowania dobroci dopasowania modelu opartą na teście chi^2. Algorytm obliczeniowy zakłada podział obserwacji z próby na podgrupy różniące się wyestymowanym na podstawie modelu prawdopodobieństwem uzyskania wartości kryterialnej zmiennej zależnej (czyli przyjęcia przez nią wartości równej 1). Następnie obliczany jest współczynnik chi^2 dla tabeli o wymiarach $g \cdot 2$, gdzie g oznacza liczbę podgrup (stanowiących kolumny tabeli), natomiast w rzędach znajdują się częstości rzeczywiste i przewidywane na podstawie modelu (Hosmer i Lemeshow, 2000). Liczba stopni swobody dla tej statystyki wynosi $g - 2$. Ponieważ test dobroci dopasowania Hosmera i Lemeshowa porównuje wartości oczekiwane na podstawie modelu z wartościami obserwowanymi, pożądanym przez badacza wynikiem jest brak istotności współczynnika chi^2. Oznacza to, że rozkład prawdopodobieństw przewidywany na podstawie modelu nie różni się istotnie od obserwowanych wyników z próby. Test ten został zaimplementowany do SPSS-a, co stanowi ogromne ułatwienie przy interpretacji wyników analizy regresji logistycznej.

Dwuczynnikowa analiza regresji logistycznej z efektem interakcyjnym

Ostatnim etapem analizy hipotetycznego eksperymentu z wpływem manipulacji eksperymentalnej i poziomu samooceny na uległość będzie włączenie do modelu regresji efektu interakcyjnego dla predyktorów. Interakcję zmiennych niezależnych w SPSS-ie wprowadza się do modelu w sposób znany chociażby z analizy wariancji, czyli przez zaznaczenie dwóch lub więcej predyktorów i kliknięcie przycisku oznaczonego symbolem > a*b >. Tym razem, zgodnie z sugestią Menarda, wybierzemy eliminację wsteczną, opartą na ilorazie wiarygodności, jako metodę wprowadzenia zmiennych do modelu. Z Opcji wybierzemy jeszcze miarę dobroci dopasowania modelu Hosmera i Lemeshowa. Przy okazji możemy zauważyć, że SPSS przyjmuje inne domyślne progi wykluczenia zmiennej z modelu dla metody selekcji postępującej (0,05) i eliminacji wstecznej (0,10). Można je oczywiście zdefiniować samodzielnie, lecz na potrzeby naszych analiz pozostawimy wartości domyślne. Po zatwierdzeniu wyborów przyciskiem OK przechodzimy do analizy raportu. Po raz kolejny widzimy statystyki dla bloku zerowego i przechodzimy do najbardziej nas interesujących wyników regresji krokowej. Krok 1 naszej analizy obejmował model pełny, czyli zawierający oba predyktory oraz ich interakcję.

Gdy wprowadzimy do modelu oba predyktory oraz ich interakcję, odpowiadające im współczynniki regresji okazują się nieistotne statystycznie. Szczególnie ważny jest dla nas fakt, że nieistotna okazała się interakcja między zmiennymi ($Z^2_{(1)} = 0{,}654$; $p = 0{,}419$). Oznacza to, że zależność między manipulacją eksperymentalną a uległością nie przebiega w odmienny sposób dla różnych

Tabela 7.
Wartości współczynników regresji dla analizy dwuczynnikowej zawierającej interakcję (baza danych: logistyczna1.sav)

		Zmienne w modelu					
		B	Błąd standardowy	Wald	*df*	Istotność	Exp(B)
Krok 1[a]	DITF	3,807	3,355	1,287	1	0,257	45,010
	SES	–0,106	0,064	2,762	1	0,097	0,899
	DITF by SES	–0,098	0,122	0,654	1	0,419	0,906
	Stała	2,404	1,876	1,642	1	0,200	11,069
Krok 2[a]	DITF	1,137	0,603	3,555	1	0,059	3,118
	SES	–0,138	0,055	6,372	1	0,012	0,871
	Stała	3,299	1,608	4,211	1	0,040	27,092

[a] – zmienne wprowadzone w kroku 1: DITF, SES, DITF * SES

wartości poziomu samooceny badanych (bądź odwrotnie) – predyktory nie pozostają w interakcji.

W kroku 2 wyłączono z modelu interakcję między predyktorami, co spowodowało wyraźny wzrost współczynników Walda i ich istotności statystycznej. Zmianę w wartościach statystyki –2 logarytm wiarygodności możemy zobaczyć w tabeli „Model – podsumowanie". Jak widzimy, wartość tej statystyki w pierwszym kroku (pełny model) jest niższa w kroku drugim, co sugeruje, że pełny model wyjaśnia więcej wariancji zmiennej zależnej niż model zredukowany, aczkolwiek zmiana ta okazała się nieistotna statycznie (co widać w tabeli „Zmienne niewłączone do modelu").

Nowym elementem naszego raportu jest tabela „Test Hosmera i Lemeshowa", zawierająca test dobroci dopasowania modelu dla obu kroków analizy.

Zarówno w pierwszym, jak i drugim kroku wartości chi^2 testu Hosmera i Lemeshowa okazały się nieistotne statystycznie. Oznacza to, że rozkład prawdopodobieństw przewidziany na podstawie wyestymowanego modelu nie różni się istotnie od wartości obserwowanych, czyli zarówno pierwszy, jak i drugi model są dobrze dopasowane do danych. Wartości bezwzględne statystyk sugerują, że model pełny (zawierający interakcję) jest nawet lepiej dopasowany do danych. Jest on jednak dla nas bezużyteczny, gdyż nie wyjaśnia więcej niż model zredukowany, a do tego wartości współczynników regresji dla modelu pełnego są nieistotne statystycznie.

W edytorze raportów znajdziemy również tabelę kontyngencji dla testu chi^2 Hosmera i Lemeshowa. Zawiera ona dokonany przez algorytm obliczeniowy podział próby na podgrupy (w naszym przypadku 9) oraz odpowiadające im liczebności oczekiwane i obserwowane. Wnikliwy czytelnik zapewne zauważył, że liczba stopni swobody dla tego testu jest rzeczywiście równa liczbie utworzonych podgrup minus 2.

Powyższy przykład obrazuje sytuację, w której testowana przez nas interakcja między predyktorami okazuje się nieistotna statystycznie. Czytelnik zapewne chciałby przeanalizować również przykład, w którym interakcja między zmiennymi niezależnymi jest statystycznie istotna. Do tego celu posłuży inna baza danych, którą można ściągnąć z serwera UMCS pod adresem: http://spoleczna.umcs.lublin.pl/pliki/logistyczna2.sav.

Tabela 8.
Wyniki testu dobroci dopasowania Hosmera i Lemeshowa

Test Hosmera i Lemeshowa

Krok	Chi-kwadrat	*df*	Istotność
1	6,781	7	0,452
2	8,509	7	0,290

Zawiera ona oprócz znanych nam zmiennych: manipulacji eksperymentalnej techniką „Drzwi zatrzaśniętych przed nosem" i zmiennej zależnej – Uległości, trzecią zmienną, a mianowicie Poziom kompetencji społecznych badanych osób (mierzony zupełnie zmyślonym narzędziem). Załóżmy, że badacz spodziewa się, iż uległość wobec technik wpływu społecznego jest modyfikowana przez poziom kompetencji społecznych badanych osób. Osoby kompetentne społecznie – jako wysoce świadome reguł rządzących ludzkim zachowaniem – miałyby być bardziej odporne na stosowanie technik wpływu społecznego (gdyż są w stanie przejrzeć grę osoby starającej się wywrzeć na nie wpływ). Można nawet założyć, że będą one oburzone próbą wpłynięcia na ich zachowanie, przez co ich uległość w warunkach stosowania technik manipulacji zachowaniem byłaby nawet niższa niż w warunkach kontrolnych. Jest to oczywiście założenie przyjęte wyłącznie na potrzeby prezentacji metody, a autor nie rości sobie pretensji do udowadniania prawdziwości tej hipotezy (przyjętej zresztą *ad hoc*).

Wprowadzamy więc do modelu regresji logistycznej zmienną zależną, oba predyktory oraz ich interakcję. Jako metodę wybieramy ponownie eliminację wsteczną opartą na ilorazie wiarygodności, a w opcjach zaznaczamy obliczanie testu dobroci dopasowania Hosmera i Lemeshowa. Uruchamiamy analizę. Tym razem analiza zakończyła się na modelu pełnym (w pierwszym kroku), obie zmienne zależne oraz ich interakcja okazały się zwiększać poziom predykcyjności modelu. Pseudo-R^2 obliczone dla modelu wyniosło 0,362 dla algorytmu Coxa i Snella oraz 0,487 dla algorytmu Nagelkerkego, co świadczy o tym, że blisko połowę wariancji zmiennej Uległość tłumaczą zmienne w modelu. Sam model jest również dobrze dopasowany do danych, wartość testu Hosmera i Lemeshowa jest nieistotna statystycznie ($chi^2_{(8)} = 12{,}239$; $p = 0{,}141$).

Przypatrzmy się wartościom współczynników regresji dla modelu. Już na pierwszy rzut oka widać, że wszystkie współczynniki regresji logistycznej okazały się istotne statystycznie. Interpretacja istotnych efektów interakcyjnych dla regresji logistycznej jest taka jak dla każdej innej analizy wielozmiennowej. Ponieważ efekt interakcyjny manipulacji eksperymentalnej i poziomu kompetencji społecznych okazał się istotny statystycznie, efekty główne tych zmiennych nie kwalifikują się do interpretacji. Jak widać, zależność między stosowaniem techniki „Drzwi zatrzaśniętych przed nosem" a Uległością jest modyfikowana przez Poziom kompetencji społecznych, dlatego nie możemy interpretować ani współczynników regresji dla

Tabela 9.
Wartości współczynników regresji dla analizy dwuczynnikowej zawierającej interakcję (baza danych: logistyczna2.sav)

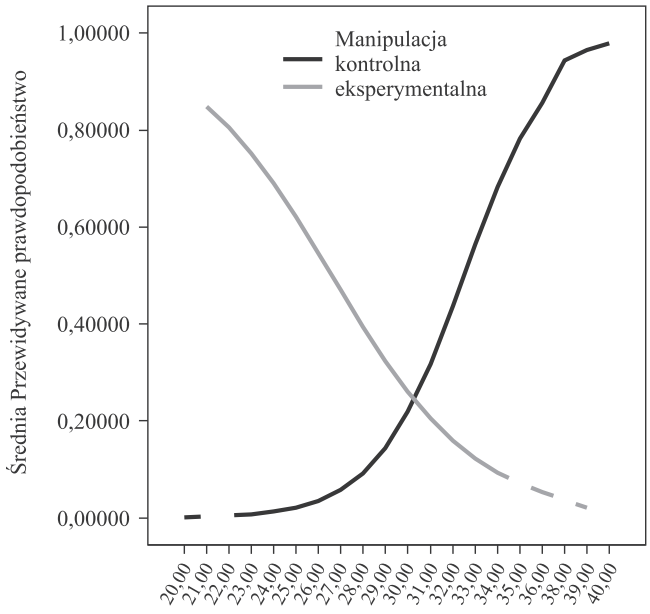
Zmienne w modelu

		B	Błąd standardowy	Wald	*df*	Istotność	Exp(B)
Krok 1[a]	DITF	24,765	6,999	12,519	1	0,000	5,695E10
	KS	0,510	0,192	7,058	1	0,008	1,665
	DITF by KS	–0,818	0,232	12,443	1	0,000	0,441
	Stała	–16,570	6,101	7,377	1	0,007	0,000

[a] – Zmienne wprowadzone w kroku 1: DITF, KS, DITF * KS

poszczególnych predyktorów, ani odpowiadających im ilorazów szans.

Analiza regresji logistycznej w programie SPSS nie umożliwia łatwego zobrazowania istotnego efektu interakcyjnego (tak jak można to zrobić w analizie wariancji). Aby to osiągnąć, musimy sami zdefiniować odpowiedni wykres. Na osi X tego wykresu powinny znaleźć się wartości naszego predyktora ilościowego, czyli poziomu kompetencji społecznych KS. Grupę eksperymentalną i kontrolną zobrazujemy jako dwie osobne linie. Pozostaje pytanie, jakie wartości powinny znaleźć się na osi Y. Nasza zmienna zależna – Uległość, przyjmuje tylko dwie wartości (0 i 1), dlatego po umieszczeniu jej na osi Y wykresu uzyskujemy mało komunikatywny „zygzak". Poza tym, chcielibyśmy uzyskać wykres obrazujący zależność opisywaną przez wyestymowany model, a nie przez dane obserwowane. W tym celu musimy stworzyć jeszcze jedną zmienną, stanowiącą przewidywane na podstawie modelu prawdopodobieństwo przyjęcia przez zmienną zależną wartości 1 dla każdej obserwacji. Na szczęście SPSS jest w stanie wygenerować taką zmienną automatycznie. W oknie definiowania regresji logistycznej klikamy przycisk Zapisz i w nowym oknie wybieramy wartości przewidywane prawdopodobieństwa. Po uruchomieniu analizy w Edytorze danych powinna pojawić się nowa zmienna (domyślnie nazwana PRE_1). Następnie z menu Wykresy wybieramy wielokrotny wykres liniowy, w którym na osi kategorii znajdzie się poziom kompetencji społecznych, zmienna manipulacji eksperymentalnej stworzy osobne linie, a nową zmienną PRE_1 wprowadzimy na oś Y jako średnią (wybierając opcję „Inna statystyka opisowa", a następnie wprowadzając tam zmienną). W ten sposób zdefiniowany wykres powinien wyglądać, jak na Rysunku 2.



Rysunek 2.
Obraz interakcji między zmienną DITF a poziomem kompetencji społecznych (KS).

Rysunek 2 jednoznacznie wyjaśnia charakter efektu interakcyjnego. Wraz ze wzrostem kompetencji społecznych zwiększa się prawdopodobieństwo spełnienia prośby eksperymentatora, ale wyłącznie w warunkach kontrolnych (pojedynczej prośby zasadniczej). W warunkach eksperymentalnych (przy stosowaniu techniki „Drzwi zatrzaśniętych przed nosem") wzrostowi kompetencji społecznych towarzyszy spadek prawdopodobieństwa spełnienia prośby zasadniczej. Wszystko wskazuje na to, że nasze dane potwierdziły zakładaną zależność między poziomem kompetencji społecznych a podatnością na wpływ społeczny.

Ilorazy szans dla efektów interakcyjnych

Pisząc o efektach interakcyjnych, do tej pory świadomie pomijano temat ilorazów szans. Wartości ilorazów szans dla zmiany jednostkowej generowane przez program SPSS są praktycznie nieinterpretowane (bo cóż znaczy jednostkowa zmiana iloczynu zmiennej DITF i KS?). Ze

względu na efekt interakcyjny sposób obliczania ilorazów szans poważnie się skomplikował. Gdy między zmiennymi zachodzi istotna statystycznie interakcja, stosowanie wzoru (8) jest błędem, gdyż wartości ilorazów szans dla dwóch grup określonych na podstawie jednej zmiennej są inne w zależności od konkretnych wartości zmiennej wchodzącej z tą pierwszą w interakcję. Dlatego wartości ilorazów szans w sytuacji interakcyjnej oblicza się zawsze dla konkretnej wartości zmiennych wchodzących w interakcje. Ogólny wzór na iloraz szans dla istotnych interakcji ma postać (Hosmer i Lemeshow, 2000):

$$(11) \quad OR_{A\times E} = e^{\sum \beta_i (x_{Ai} - x_{Bi}) + \sum \delta_j W_j (x_{Aj} - x_{Bj})}$$

gdzie: x_{Ai} oznacza wartość i-tej zmiennej niezależnej w grupie A
x_{Bi} oznacza wartość i-tej zmiennej niezależnej w grupie B
β_i oznacza współczynnik regresji dla i-tej zmiennej niezależnej
δ_j oznacza współczynnik regresji dla j-tej zmiennej interakcyjnej
W_j oznacza konkretną wartość zmiennej, która wchodzi w j-tą interakcję
x_{Aj} oznacza wartość j-tej zmiennej interakcyjnej w grupie A
x_{Bj} oznacza wartość j-tej zmiennej interakcyjnej w grupie B

Powyższy wzór jest bardzo uciążliwy, gdyż wymaga podstawiania wartości zmiennych interakcyjnych, będących najczęściej iloczynem zmiennych wchodzących w te interakcje. Mimo, że pozwala obliczyć iloraz szans dla dowolnie zdefiniowanych grup (ze względu na wszystkie możliwe wartości wszystkich zmiennych niezależnych), jest najczęściej mało użyteczny.

Znacznie bardziej przyjazny obliczeniowo (i interpretacyjnie) jest tzw. skorygowany iloraz szans (Kleinbaum i Klein, 2002). Zakłada on obliczenie ilorazu szans dla dwóch grup zdefiniowanych na podstawie jednej zmiennej, przy kontroli wpływu pozostałych zmiennych. Łatwe obliczenie takiego ilorazu szans umożliwia fakt, że gdy wprowadzimy do regresji logistycznej w jednym modelu kilka zmiennych niezależnych niepozostających ze sobą w interakcji, ich współczynniki regresji zostają tak skorygowane, by uwzględniały wpływ pozostałych zmiennych (była o tym mowa przy analizie dwuczynnikowej bez interakcji). Skorygowany iloraz szans dla modelu bez interakcji ma postać:

$$(12) \quad OR_{A\times B} = e^{\beta (x_A - x_B)}$$

gdzie: β oznacza współczynnik regresji dla zmiennej niezależnej różnicującej grupę A i B
x_A oznacza wartość różnicującej zmiennej niezależnej w grupie A
x_B oznacza wartość różnicującej zmiennej niezależnej w grupie B

Ten sam wzór dla modelu zawierającego interakcję ma postać:

$$(13) \quad OR_{A\times B} = e^{\beta (x_A - x_B) + \sum \delta_j W_j}$$

gdzie: β oznacza współczynnik regresji dla zmiennej niezależnej różnicującej grupę A i B
x_A oznacza wartość różnicującej zmiennej niezależnej w grupie A
x_B oznacza wartość różnicującej zmiennej niezależnej w grupie B
δ_j oznacza współczynnik regresji dla j-tej zmiennej interakcyjnej (wchodzącej z w interakcję ze zmienną różnicującą)
W_j konkretna wartość zmiennej wchodzącej w j-tą interakcję ze zmienną różnicującą

Zdając sobie sprawę, że powyższe wzory mogą być mocno abstrakcyjne, korzystając ze wzoru (13), spróbujmy obliczyć skorygowany iloraz szans, porównujący szansę na uzyskanie uległości w grupie eksperymentalnej z grupą kontrolną z ostatniego przykładu. Za x_A podstawimy 1 (gdyż wartość zmiennej DITF dla grupy eksperymentalnej wynosi właśnie 1), za x_B podstawimy 0 (wartość zmiennej DITF dla grupy kontrolnej), a za β wartość współczynnika regresji dla zmiennej DITF (24,765). Ponieważ mamy tylko jeden efekt interakcyjny, za δ_j podstawimy współczynnik regresji dla składnika interakcyjnego (–0,818). Chcąc zobrazować interakcję między zmiennymi, musimy obliczyć co najmniej dwa ilorazy szans porównujące te dwie grupy przy co najmniej dwóch różnych wartościach zmiennej KS (poziom kompetencji społecznych). Ponieważ dla naszych zmyślonych przecież danych uzyskaliśmy dosyć strome postaci funkcji logarytmicznej, za dwa punkty odniesienia przyjmiemy wartość pierwszego i trzeciego kwartyla zmiennej KS (czyli $Q_1 = 25$ i $Q_3 = 32$). Podstawiając dane do wzoru (13), uzyskujemy następujące ilorazy szans

$$OR_{A\times B\ (25)} = e^{24{,}765(1-0)+(-0{,}818)25} = e^{4{,}315} \approx 74{,}81$$

Dla osób o niższych kompetencjach społecznych (na poziomie pierwszego kwartyla) należy spodziewać się, że szansa uzyskania uległości w grupie eksperymentalnej będzie prawie 75-krotnie większa niż w grupie kontrolnej.

Stąd wniosek, że wobec osób mało kompetentnych społecznie warto stosować techniki wpływu społecznego.

$$OR_{A\times B\ (32)} = e^{24,765(1-0)+(-0,818)32} = e^{-1,411} \approx 0,24$$

Natomiast w grupie osób o wyższych kompetencjach społecznych (na poziomie trzeciego kwartyla) szansa na uzyskanie uległości w warunkach stosowania techniki „Drzwi zatrzaśniętych przed nosem" jest prawie 4-krotnie niższa niż w sytuacji stosowania niczym niepoprzedzonej prośby. Oczywiste jest więc, że stosowanie technik wpływu społecznego wobec osób kompetentnych społecznie może okazać się nie tylko nieskuteczne, ale nawet przeciwskuteczne (co jest oczywiście pobożnym życzeniem, zważywszy chociażby na fakt, co było motywacją Roberta Cialdiniego do napisania książki *Wywieranie wpływu na ludzi*).

Zamiast podsumowania, czyli o czym nie napisano

Zawarcie w jednym artykule najważniejszych informacji na temat tak złożonej analizy statystycznej, jaką jest regresja logistyczna, okazało się nie lada wyzwaniem. Ograniczono do minimum prezentację wzorów i procedur obliczeniowych, zrezygnowano z opisu przekształcenia logitowego i jego związku z ilorazem szans. Autor pominął zupełnie drugie zastosowanie regresji logistycznej, jakim jest (oprócz testowania hipotez) obliczanie przedziałów ufności dla współczynników regresji i odpowiadających im ilorazów szans. Nie było to jednak najistotniejsze przy stosowaniu tej metody do analizy danych eksperymentalnych.

Z uwagi na objętość artykułu nie zawarto w nim procedury stosowania regresji logistycznej dla więcej niż dwukategorialnych predyktorów jakościowych. Jest to oczywiście możliwe, aczkolwiek wymaga kodowania takich zmiennych na $k - 1$ wektorów w sposób podobny, w jaki włącza się te zmienne do regresji liniowej. Sposób wprowadzania dychotomicznych predyktorów przedstawiony w tym opracowaniu jest przykładem najprostszego kodowania zmiennych nominalnych, tzw. *dummy coding,* czyli kodowania zero-jedynkowego. W przypadku schematów badawczych zawierających kilka wielokategorialnych nominalnych zmiennych niezależnych bardziej użyteczną analizą niż regresja logistyczna staje się analiza log-liniowa.

W niniejszym artykule przedstawiono najbardziej klasyczną wersję regresji logistycznej. Istnieje również forma wielomianowa, dla wielokategorialnych jakościowych zmiennych zależnych, jak również odmiana dla zmiennych porządkowych (tzw. PLUM). Jeszcze inny model stosuje się dla danych skorelowanych. Są one rozwinięciem podstawowego modelu dwumianowego i dla ich zrozumienia niezbędne jest zapoznanie się z omówioną tutaj wersją klasyczną.

LITERATURA CYTOWANA

Aronson, E., Wilson, T. D., Akert, R. M. (1997). *Psychologia społeczna. Serce i umysł.* Poznań: Zysk i S-ka.

Bronsztajn, I. N., Siemiendiajew, K. A. (1970). *Matematyka. Poradnik encyklopedyczny*. Warszawa: Państwowe Wydawnictwo Naukowe.

D'Agostino, R. B. (1971). A second look at analysis of variance on dichotomous data. *Journal of Educational Measurement, 8(4),* 327–333.

Doliński, D. (2000). *Psychologia wpływu społecznego*. Wrocław: Towarzystwo Przyjaciół Ossolineum.

Ferguson, G. A., Takane, Y. (1999). *Analiza statystyczna w psychologii i pedagogice*. Warszawa: Państwowe Wydawnictwo Naukowe.

Field, A. (2005). *Discovering statistics using SPSS* (wyd. 2). London: Sage Publishing.

Hosmer, D. W., Lemeshow, S. (2000). *Applied logistic regre*ssion (wyd. 2). New York: Wiley & Sons.

Kleinbaum, D. G., Klein, M. (2002). *Logistic regression – a self-learning text* (wyd. 2). New York: Springer.

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, *4(2)*, 263–269.

Menard, S. (1995). *Applied logistic regression analysis*. London: Sage University Papers.

Stanisz, A. (2000). *Przystępny kurs statystyki z wykorzystaniem programu STATISTICA PL na przykładach z medycyny* (t. 2). Kraków: Wydawnictwo StatSoft Polska.

SPSS 17.0 PL, podręcznik elektroniczny.

PRZYPISY

1. Współczynniki B generowane przez program SPSS to tak naprawdę współczynniki beta (język programu nie zawiera znaków greckich, więc jego twórcy uprościli zapis, zastępując go znakiem B).

Application of logistic regression in experimental research

Barnaba Danieluk

Institute of Psychology, Maria Curie-Skłodowska University, Lublin

Abstract

In experimental practice we often face the situation where the measured dependent variable takes one of two values only: 0 – lack of the measured characteristic or 1 – observation of the measured characteristic (behavior, consent to something, displaying an attitude or an opinion etc.). Both the general linear model as well as the linear regression analysis cannot be applied to dichotomous, nominal dependent variables. In such cases we are forced to use the non-linear analysis. Logistic regression is the model used for this type of dependent variables. This article presents application of the binomial logistic regression in experimental research. It explains specification and interpretation of typical logistic regression coefficients such as odds ratio, Wald coefficients, likelihood ratios. It presents the estimation procedure of the model parameters with *maximum likelihood procedure* and the *Hosmer*-Lemeshow *goodness* of *fit* test. Introduced were simple sample analyses (with nominal and quantitative predictors), a two-factor analysis as well as a two-factor analysis with interaction effect. The number of formulas and algebraic transformations were cut to the necessary minimum and the shown sample analysis and their interpretation were conducted step by step with the SPSS Statistics Pack version 17.0 PL.

Key words: logistic regression, binomial logistic regression, odds ratio, likelihood ratio, maximum likelihood method, Wald coefficient, SPSS Statistics

Złożono do druku: 22.02.2010
Złożono poprawiony tekst: 9.05.2010
Zaakceptowano do druku: 16.05.2010