Przemysław KOROHODA [1], Monika MIKLASZEWSKA [2], Jacek Antoni PIETRZYK [2]
[1] AGH AKADEMIA GÓRNICZO-HUTNICZA, KATEDRA ELEKTRONIKI, Al. Mickiewicza 30, 30-059 Kraków
[2] KLINIKA NEFROLOGII DZIECIĘCEJ KATEDRY PEDIATRII WYDZIAŁU LEKARSKIEGO UJ W KRAKOWIE, ul. Wielicka 265, 30-663 Kraków

Graficzna interpretacja wyników regresji logistycznej w badaniu klinicznym dla dwustanowej zmiennej zależnej

Dr inż. Przemysław KOROHODA

Absolwent wydziału Elektroniki, Automatyki i Elektroniki Akademii Górniczo-Hutniczej w Krakowie. Obecnie pracuje jako adiunkt w Katedrze Elektroniki AGH. W ostatnim okresie jego zainteresowania naukowe dotyczą modelowania hemodializy, analizy bioimpedancji wieloczęstotliwościowej oraz uogólnienia teorii filtracji liniowej. Prowadzi zajęcia dydaktyczne z zakresu przetwarzania sygnałów i obrazów cyfrowych.

e-mail: korohoda@agh.edu.pl

Dr n. med. Monika MIKLASZEWSKA

Absolwent Collegium Medicum Uniwersytetu Jagiellońskiego w Krakowie. Obecnie zatrudniona w Klinice Nefrologii Dziecięcej UJ CM na stanowisku adiunkta. W pracy naukowej i klinicznej zajmuje się głównie problematyką ostrego uszkodzenia nerek, markerami predykcyjnymi AKI, przewlekłą chorobą nerek, powikłaniami układu sercowo – naczyniowego, badaniami urodynamicznymi oraz obrazowaniem USG.

e-mail: mmiklasz@mp.pl

Streszczenie

W artykule zaproponowano graficzną formę prezentacji wyników regresji logistycznej, zastosowanej dla dwustanowej zmiennej zależnej i dwóch numerycznych zmiennych pomiarowych. Przed wyznaczeniem współczynników regresji jedna z tych zmiennych została poddana transformacji z wykorzystaniem logarytmu o podstawie 2. Opisano sposób przeprowadzenia odpowiednich obliczeń i zaprezentowano przykładowe wyniki, bazując na wynikach badania klinicznego (dla N=47 pacjentów) stężeń interleukiny 6 w surowicy dzieci poddawanych zabiegom kardiochirurgicznym w krążeniu pozaustrojowym, narażonych na ryzyko wystąpienia ostrego uszkodzenia nerek (OUSzN). Metoda pozwala na szybką identyfikację ryzyka i podjęcie właściwych działań terapeutycznych, w tym – terapii nerkozastępczej.

Słowa kluczowe: regresja logistyczna, logit, szansa, nomogram, interleukina 6 w surowicy, ostre uszkodzenie nerek.

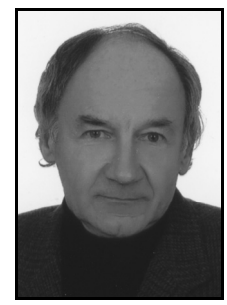
Graphical interpretation of logistic regression results for clinical investigation with binomial output variable

Abstract

In the paper there are proposed nomograms being graphical presentation of logistic regression results. Such graphs depict the probability and odds functions of the considered feature in a readable form. The investigated particular case is suited for the binomial output variable and two numerical input variables. The suggested presentation form has been selected with focus on clinical personnel accustomed to such form of data presentation. After introductory explanation of the logistic regression essentials with emphasis on the Gaussian-based distribution, there are presented assumption and resulting - alternative model formulation supported by the simulation experiment results, basing on the clinically collected data from 47 patients suspected to experience acute kidney injury (AKI) diagnosed accordingly to interleukin 6 serum concentration. The interpretation of unit selection for each input variable is shortly discussed along with the consequences of possible logarithmic transformation. The presented method allows for either quick, graphical, bedside risk identification and mapping of lifethreating condition or early introduction of renal replacement therapy.

Keywords: logistic regression, logit, odds, nomogram., urine interleukin 6 concentration, acute kidney injury.

Prof. dr hab. med. Jacek Antoni PIETRZYK



Specjalista chorób dzieci, kierownik Kliniki Nefrologii Dziecięcej i Zakładu Dializ w Polsko-Amerykańskim Instytucie Pediatrii Collegium Medium UJ w Krakowie. Od 1987 roku współpracuje z Katedrą Elektroniki AGH, realizując 7 kolejnych projektów naukowo-badawczych KBN związanych z modelowaniem kinetycznym i optymalizacją dializy, diagnostyką dostępu naczyniowego, biozgodnością i efektywnością reutylizacji dializatorów oraz zastosowaniem analizy bioimpedancji wieloczęstotliwościowej.

e-mail: jacek.a.pietrzyk@gmail.com

1. Wprowadzenie

Klasyczne równanie regresji liniowej, modelujące za pomocą równania liniowego wiążącego jedną lub więcej, na przykład N, zmiennych wyjaśniających x z pojedynczą zazwyczaj zmienną zależną v

$$v = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_N x_N , \tag{1}$$

stanowi jedno z podstawowych narzędzi opisujących zależności między zmiennymi numerycznymi w badaniach statystycznych realizowanych dla danych medycznych [5, 7, 8, 9]. Na podstawie zbioru wyników pomiarowych wyznaczane są współczynniki równania, przy założeniu minimalizowania błędu modelu określonego jako błąd średniokwadratowy. Zadanie to można rozwiązać wykorzystując zapis macierzowy, gdzie rozwiązanie układu równań, których liczba znacznie przewyższa liczbę wyznaczanych współczynników, czyli niewiadomych, można zrealizować wykorzystując macierz pseudoodwrotną [6]. Można wykazać, że rozwiązanie takie zapewnia minimalizację podanego typu błędu.

Stosunkowo często zachodzi jednak konieczność zaproponowania modelu dla zależności, gdy zmienna zależna jest dyskretna, w szczególności dwustanowa. Przykładowo dla konkretnego pacjenta zachodzi albo występowanie określonego objawu/choroby, albo brak tego objawu/choroby. Jest to cecha, która można przyjmować umowne wartości 0 lub 1. W przypadku, gdy zbiór rozważanych pacjentów posiada ponadto pewne przypisane cechy, opisane numeryczne, jak na przykład: wzrost, masę ciała, wiek, dawkę przyjętego leku, czy stężenie wybranej substancji markerowej, w doniesieniach literaturowych [1, 2, 3] stosuje się model regresji logistycznej [5, 7, 8, 9], w którym zmienną zależną jest prawdopodobieństwo wystąpienia badanej cechy u pacjenta o określonym zestawie wartości zmierzonych parametrów niezależnych

$$\ln \frac{p}{1-p} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_N x_N , \tag{2}$$

Wyznaczanie współczynników b realizowane jest w tym przypadku poprzez maksymalizowanie wartości funkcji wiarygodności. Zadanie to może być realizowane za pomocą specjalizowanych pakietów, jak na przykład Statistica [11], czy Matlab [10]. Istotnym założeniem, którego spełnienia w praktyce nie zawsze możemy być pewni, jest by zmienne wyjaśniające posiadały rozkłady normalne. W przypadku, gdy potrafimy zaproponować odpowiednią transformację, zmieniającą rozkład danej zmiennej w rozkład normalny, to należy ją zastosować. Taką transformacją może być na przykład funkcja logarytmu.

Model regresji logistycznej oferuje dodatkowe możliwości wnioskowania, które w ostatnich latach wydają się być szczególnie chętnie stosowane w analizie medycznych danych pomiarowych [1, 2, 3]. Na podstawie współczynników b (2) wyznaczany

jest parametr ilorazu szans (*ang. Odds Ratio* - *OR*), który dla pojedynczej zmiennej x_k wyraża się zależnością

$$OR_k = e^{b_k}, \tag{3}$$

Z punktu widzenia przydatności klinicznej bardzo istotne jest prawidłowe odczytanie informacji zawartej w tym parametrze.

2. Interpretacja współczynników regresji

Należy zauważyć, że model, typowo opisywany jedynie równaniem (2), można również zinterpretować nieco inaczej. Stałość parametru *OR* dla wszystkich wartości zmiennej niezależnej może być uzyskana jedynie, gdy założymy, że populacja może być podzielona na dwie grupy – odpowiednio z wartością cechy 0 oraz 1 – przy czym każda z tych dwóch części jest opisana przez identyczny rozkład normalny, różniący się jedynie wartością średnią, czyli $\mu_0 \neq \mu_1$, natomiast odchylenia standardowe pozostają takie same, $\sigma_0 = \sigma_1 = \sigma$. Oznacza to, iż występowanie, lub nie, danej cechy wiąże się jedynie z przesunięciem całego rozkładu o pewną stałą i wtedy zachodzi następująca zależność

$$b_k = \frac{\mu_1 - \mu_0}{\sigma^2}, \tag{4}$$

W takim przypadku iloraz szans oznacza proporcję dwóch wartości szansy wystąpienia cechy, wyliczonej dla x_k+1 oraz dla x_k.

Zmienna x_k wyrażona jest w określonych jednostkach. Szansa wyraża proporcję prawdopodobieństwa wystąpienia cechy do prawdopodobieństwa jej niewystąpienia. Przykładowo, gdy x_k jest stężeniem w g/mL oraz OR_k=1,5, oznacza to, że zwiększenie stężenia x_k o 1 g/mL jest powiązane z 1,5-krotnie większą szansą wystąpienia cechy. Gdyby przed wyznaczeniem współczynników regresji podzielono wartości zmiennej x_k przez 100, to wyznaczony parametr *OR* wskazywałby na zmianę szansy przy każdym wzroście stężenie o 100 g/mL. Daje to możliwość wyboru jednostki wygodnej w ocenie klinicznej. W przypadku, gdyby x_k było zmienną otrzymaną przez wyznaczenie logarytmu o podstawie 2 ze stężenia, to parametr *OR*=1,5 oznaczałby 1,5-krotny wzrost szansy przy podwojeniu stężenia (podstawa logarytmu). W tym przypadku jednostki pierwotne nie mają znaczenia. Jak widać, zastosowanie logarytmu naturalnego, oznaczałoby konieczność interpretowania *e*-krotnych (*e*=2,73…) zmian zmiennej pierwotnej.

Wykorzystując model regresji logistyczne dla wielu zmiennych wyjaśniających, parametr *OR* można także wyznaczać łącznie dla kilku wybranych zmiennych

$$OR_{(k,m,n)} = e^{b_k} e^{bm} e^{b_n} = e^{b_k + b_m + b_n}, \tag{5}$$

Wyraża on oszacowanie wzrostu szansy, przy jednoczesnym wzroście o jednostkę wszystkich wykorzystanych zmiennych. W tym przypadku x_k, x_m oraz x_n. Każda ze zmiennych może mieć swoją własną, odpowiednio dobraną jednostkę a wybrane zmienne mogą być wynikiem np. logarytmowania zmiennych pierwotnych. Jednak interpretacja kliniczna takiego wyniku nie jest już tak oczywista, jak w przypadku *OR* dla pojedynczej zmiennej, zatem jego przydatność może być w praktyce mocno ograniczona. Rozważając jednoczesną interpretację wyników otrzymanych dla dwóch zmiennych pochodzących z pomiaru w ramach rzeczywistego eksperymentu klinicznego, stwierdzono, iż wskazane jest określanie nie tylko przyrostu szans, ale także samych wartości szans, i to nie tylko przy jednoczesnych jednostkowych przyrostach obu zmiennych, ale dla bardziej realnych klinicznie sytuacji. Ze względu na możliwość wygodnej reprezentacji graficznej, podejście takie staje się szczególnie wygodne dla dwóch zmiennych wyjaśniających. Dla uproszczenia w dalszych rozważaniach zmienne te będą oznaczone jako *x* i *y*.

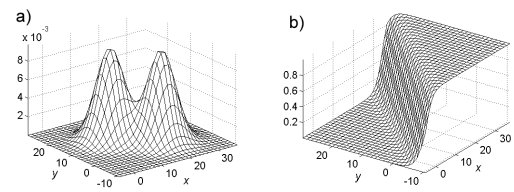
3. Eksperyment symulacyjny

W celu zilustrowania zaproponowanej interpretacji opisanego modelu, przeprowadzono obliczenia symulujące idealny eksperyment kliniczny. W pakiecie Matlab wygenerowano $N=2^{14}$ zestawów danych, przyjmując iż połowa z nich odpowiada wartości cechy 0, a połowa 1. Rolę zmiennych wyjaśniających pełniły dwie zmienne numeryczne o rozkładach normalnych i odchyleniu standardowym σ=4, oraz wartościach średnich: μ_{x0}=10, μ_{x1}=20, μ_{y0}=15, μ_{y1}=5.

Funkcję prawdopodobieństwa wystąpienia cechy, w zależności od wartości zmiennych *x* oraz *y* wyznaczono na podstawie zależności

$$p(x,y) = \frac{e^{z(x,y)}}{1+e^{z(x,y)}}; \quad z(x,y) = b_0 + b_1 x + b_2 y, \tag{6}$$

Na rys. 1 pokazano sumaryczny rozkład tak opisanych danych oraz trójwymiarowy wykres prawdopodobieństwa, wynikający z wyliczenia parametrów dla regresji logistycznej.



Rys. 1. Dla przeprowadzonego eksperymentu: a) suma obu (dla cechy 0 oraz 1) funkcji łącznej gęstości rozkładów prawdopodobieństwa dla zmiennych numerycznych, b) wykres funkcji prawdopodobieństwa wystąpienia cechy, wyznaczonej w wyniku zastosowania regresji logistycznej

Fig. 1. For the conducted experiment: a) sum of both (for feature being 0 and 1) joint probability density functions for the numerical variables, b) graph of the probability function for the binomial feature, obtained from logistic regression analysis

Z wykresu na rys. 1b odczytać można obszary, gdzie prawdopodobieństwo wystąpienia cechy jest bliskie 1 oraz gdzie jest bliskie 0. Dokładniejszy odczyt nie jest jednak możliwy. W tabeli 1 zawarto dane pośrednie oraz wartości ilorazów szans.

Tab. 1. Wyznaczone współczynniki regresji oraz ilorazy szans

Tab. 1. Computed regression coefficients and odds ratios

k	0	1 (x)	2 (y)
b_k	-3,2375	0,6208	-0,6119
OR_k	-	1,8604	0,5423

Natomiast łączny iloraz szans wynosi: $OR_{(1,2)}$= 1,0089.

Wyniki powyższe należy interpretować następująco. Wzrost zmiennej *x* o 1 powoduje wzrost szansy o 86%, wzrost jednostkowy zmiennej *y* wiąże się ze zmniejszeniem szansy o 66%, natomiast jednoczesna zmiana jednostkowa obu zmiennych nie powoduje praktycznie żadnej zmiany wartości szansy.

4. Wyniki kliniczne

Dla *N*=47 pacjentów pediatrycznych z wrodzoną wadą serca, którzy poddani byli zabiegowi kardiochirurgicznemu w krążeniu pozaustrojowym, w drugiej godzinie po zabiegu zbadano stężenie w surowicy interleukiny 6 (IL6) [2, 3], a po upływie doby dokonano oceny wystąpienia ostrego uszkodzenia nerek, zgodnie z kryteriami według skali *pediatric* pRIFLE [1]. W wyniku tej procedury otrzymano próbę *N*0=28 dzieci, u których cecha (ostre uszkodzenie nerek) nie wystąpiła oraz *N*1=19 dzieci, dla których stwierdzono wystąpienie cechy. Liczność próby umożliwiła przeprowadzenie analizy regresji logistycznej dla dwóch zmiennych [4], przy czym po wstępnych analizach oprócz IL6, wyrażanej w pg/mL, jako drugą zmienną wytypowano masę ciała w kg. Za jednostkę zmian stężeń IL6 przyjęto 100 pg/mL, natomiast dla masy ciała konieczne było zastosowanie transformacji logarytmicznej, wybrano zatem logarytm o podstawie 2. W tabeli 2 zebrano wartości pośrednie oraz wyliczone ilorazy szans.

Tab. 2. Wyznaczone współczynniki regresji oraz ilorazy szans
Tab. 2. Computed regression coefficients and odds ratios

k	0	1 (x)	2 (y)
b_k	1,4545	0,6811	-1,1204
OR_k	-	1,9760	0,3261

Wyniki zawarte w tabeli 2 wskazują, że każdy przyrost o 100 pg/mL stężenia IL6 wiąże się ze wzrostem szansy wystąpienia ostrego uszkodzenia nerek o 98%, natomiast każde podwojenie masy ciała zmniejsza tę szansę o 67%. Łączny iloraz szans wynosi w tym przypadku: $OR_{(1,2)}$= 0,6445. Wynik ten oznacza, że jednoczesne zwiększenie stężenia IL6 o 100pg/mL oraz podwojenie masy ciała dają łącznie spadek szansy ostrego uszkodzenia nerek o 35%. Łatwo zauważyć, że bez odpowiedniego przeliczenia jest to informacja mało użyteczna w warunkach klinicznych. Dlatego też opracowano graficzną metodę prezentującą wyniki wyrażone powyższymi liczbami w postaci odpowiedniego nomogramu. Podczas tworzenia takiego wykresu możliwe było dodatkowo sprowadzenie wyników do jednostek pierwotnych – np. ponowne wyrażenie masy ciała w kg, a nie poprzez wartości logarytmu.

Po wyznaczeniu współczynników regresji wyznaczono wartości pomocniczej funkcji $z(x,y)$, zastosowanej w (6)

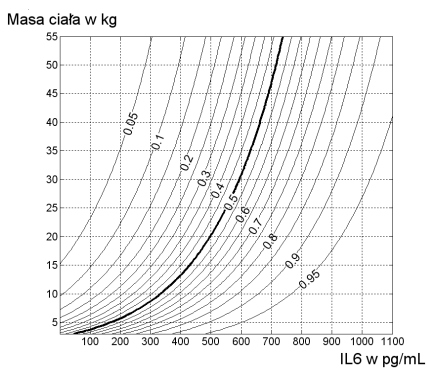
$$z(x,y) = b_0 + b_1 \frac{x}{100} + b_2 \log_2(y), \tag{7}$$

gdzie x oznacza steżenie IL6 w pg/mL, natomiast y - masę ciała w kg.

Bazując na założeniach modelu wyznaczono zależność dla szansy jako funkcji obu zmiennych

$$s(x,y) = \frac{p(x,y)}{1-p(x,y)}, \tag{8}$$

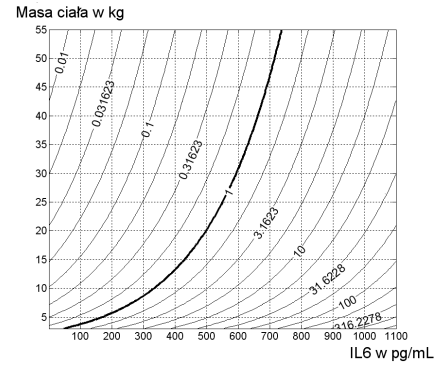
Funkcję prawdopodobieństwa, wyliczoną według zależności (6) z podstawieniem (7), można przedstawić w formie trójwymiarowego wykresu, jak na rys. 1b. Jednak w celu odczytania wartości prawdopodobieństwa dla konkretnych danych, w podręcznikach zalecane jest podstawienie tych danych do wzoru (6) i wyliczenie pojedynczej wartości prawdopodobieństwa.



Rys. 2. Nomogram funkcji prawdopodobieństwa (6) wynikającej z rozwiązania zadania regresji logistycznej dla danych klinicznych
Fig. 2. Nomogram for the probability function (6) being the result of logistic regression task solution for the clinically obtained data

Zastąpienie trójwymiarowego wykresu przez wykres poziomicowy, pokazany na rys. 2, umożliwia nie tylko dostatecznie precyzyjny odczyt, ale także szybką ocenę, na ile zmieni się wartość prawdopodobieństwa, przy założonym marginesie zmian parametrów niezależnych.

Przeliczenie (8) wyznacza funkcję szansy, przy czym w tym przypadku wykres trójwymiarowy staje się jeszcze bardziej nieczytelny, z powodu faktu, iż wartości tej funkcji obejmują zwykle kilka rzędów wielkości. Sporządzenie odpowiedniego nomogramu, jak na rys. 3, z wykładniczo narastającymi wartościami poziomów, umożliwia natychmiastowe, odpowiednio dokładne określenie wartości szansy dla danego zestawu parametrów.



Rys. 3. Nomogram funkcji szansy (8) wynikającej z rozwiązania zadania regresji logistycznej dla danych klinicznych
Fig. 3. Nomogram for the odds function (6) being the result of logistic regression task solution for the clinically obtained data

Pogrubiona linia na rys. 3 odpowiada pogrubionej linii z rys. 2.

5. Wnioski

Proponowana forma przedstawiania wyników regresji logistycznej jest stosunkowo prosta i znana z wielu innych dziedzin, jednak pomimo tego nie została dotąd zastosowana w dostępnych podręcznikach ani publikacjach opisujących wyniki badań eksperymentalnych. Opracowane nomogramy przedstawiające linie stałych wartości szansy oraz ryzyka wystąpienia OUSzN, wskazujące na przedziały tych wartości w zależności od wartości stężenia IL6 w surowicy w odniesieniu do masy ciała mogą stanowić cenną pomoc w przyłóżkowej ocenie ryzyka wystąpienia ostrego uszkodzenia nerek i wczesnych interwencji terapeutycznych, w tym – terapii nerkozastępczej.

6. Literatura

[1] Akcan-Arikan A., Zappitelli M., Loftis L. L., Washburn K. K., Jefferson L. S., Goldstein S. L.: Modified RIFLE criteria in critically ill children with acute kidney injury. Kidney International, 71, 1028–1035, 2007.
[2] Li Y., Fu C., Zhou X., Xiao Z., Zhu X., Jin M., Li X., Feng X.: Urine interleukin-18 and cystatin-C as biomarkers of acute kidney injury in critically ill neonates. Pediatric Nephrology, DOI 10.1007/s00467-011-2072-x, to be published in 2012.
[3] Liu K. D., Altmann C., Smits G., Krawczeski C. D., Edelstein C. L., Devarajan P., Faubel S.: Serum Interleukin-6 and interleukin-8 are early biomarkers of acute kidney injury and predict prolonged mechanical ventilation in children undergoing cardiac surgery: a case-control study. Critical Care, 13, R104, 2009.
[4] Peduzzi P., Concato J., Kemper E., Holford T.R., Feinstein A.R.: A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49, 1996.
[5] Petrie A., Sabin C.: Statystyka medyczna w zarysie. Wydawnictwo Lekarskie PZWL, Warszawa 2006.
[6] Press W.H., Teukolsky S.A. Vetterling W.T., Flannery B.P.: Numerical Recipes in C. Wyd. 2., Cambridge University Press, Cambridge 1992.
[7] Stanisz A. (red.): Biostatystyka, podręcznik dla studentów medycyny. Wyd. UJ, Kraków 2005.
[8] Stanisz A.: Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Wyd. 3., Statsoft Polska, Kraków 2006.
[9] Watała C.: Biostatystyka – wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych. Alfa-medica Press, Bielsko-Biała 2002.
[10] MatLab, version 7.5, Mathworks, www.mathworks.com.
[11] Statistica (data analysis software system), Version 10, StatSoft, Inc., www.statsoft.com.

otrzymano / received: 16.01.2012
przyjęto do druku / accepted: 02.03.2012 *artykuł recenzowany / revised paper*