

Regresja logistyczna

Podstawy statystyki dla prowadzących badania naukowe

Odcinek 30: Regresja logistyczna

mgr Andrzej Stanisław

z Zakładu Biostatystyki i Informatyki Medycznej Collegium Medicum UJ w Krakowie

Data utworzenia: 23.04.2002

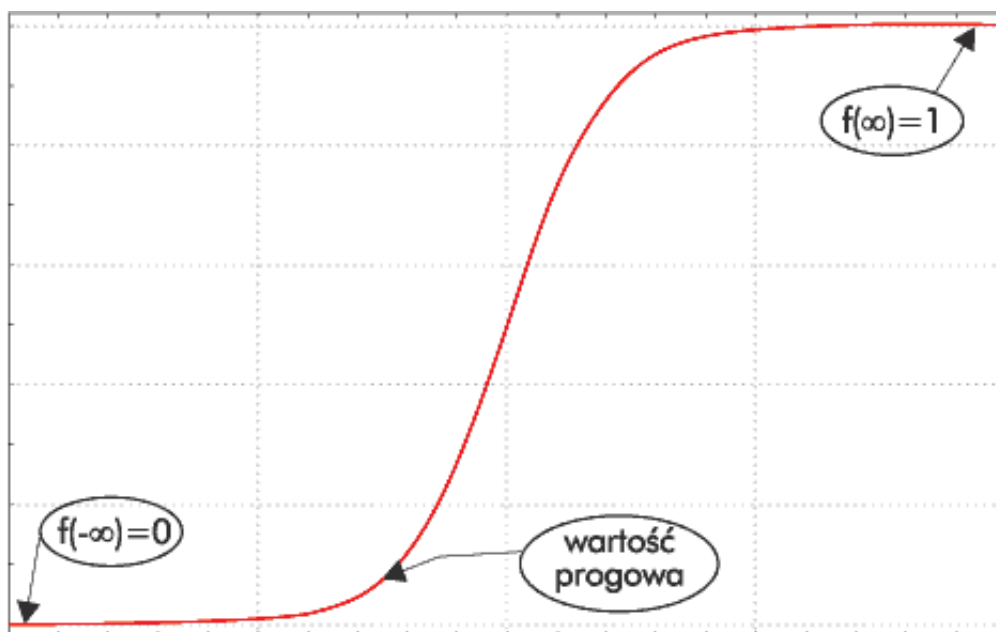
Ostatnia modyfikacja: 30.04.2007

Opublikowano w Medycyna Praktyczna 2001/10

W poprzednim odcinku opisaliśmy modele linearyzowalne regresji nieliniowej, tj. takie, które poprzez odpowiednie przekształcenia liniowe zmiennych lub współczynników możemy sprowadzić do postaci liniowej. Poniżej przedstawię inny rodzaj modelu nieliniowego - regresję logistyczną, bardzo popularną w naukach medycznych. W badaniach medycznych mamy często do czynienia ze zmienną zależną typu dychotomicznego, na przykład: 1 - przeżycie, 0 - zgon; 1 - występuje dany objaw (np. hiperglikemia, gorączka), 0 - nie ma danego objawu; albo 1 - występuje nowotwór, 0 - nie ma nowotworu. Wówczas można postawić pytanie, która ze zmiennych niezależnych wpływa istotnie na przeżycie lub wystąpienie objawu. Szukamy oczywiście podobnego do funkcji regresji powiązania prawdopodobieństwa wystąpienia danego objawu z grupą zmiennych niezależnych, takich jak wiek, płeć, stężenie cholesterolu itd.

W takich sytuacjach nie możemy wykorzystać regresji wielokrotnej (p. Regresja wielokrotna - cz. I oraz Regresja wielokrotna - cz. II). Wartości estymatora takiej funkcji regresji mogą bowiem przyjmować wartości mniejsze od zera lub większe od jedności, które w naszym przypadku nie mają interpretacyjnego sensu. Doskonałym narzędziem do tego typu zagadnień jest właśnie regresja logistyczna. Dodatkową zaletą regresji logistycznej jest to, że analiza i interpretacja wyników są podobne jak w poznanych wcześniej metodach klasycznej regresji. Ogólnie mówiąc, regresja logistyczna jest pewnym matematycznym modelem, którego możemy użyć w celu opisanie wpływu kilku zmiennych $x_1, x_2, \dots, x_k$ (zarówno ilościowych, jak i jakościowych) na dychotomiczną zmienną y .

Regresję logistyczną, cieszącą się dużym uznaniem w psychologii i medycynie, wprowadzono już w XIX wieku. Pionierami byli Verhulst i Pearl, którzy opracowali postać krzywej logistycznej i zastosowali ją w praktyce; jednakże pełny model w postaci stosowanej dzisiaj po raz pierwszy podał i zastosował w 1972 roku Finney. Co leży u podstaw tej popularności? Przede wszystkim funkcja logistyczna, której wykres widoczny jest na rysunku 1.



Rys. 1. Postać funkcji logistycznej

Jak widać, przyjmuje ona wartości między 0 a 1. Zakłada się ponadto, że jej wartość w $-\infty$ jest równa 0

$(f[-\infty] = 0)$, a dla $z = +\infty$ $f(z)$ jest równa 1.

Jak zmienia się wartość funkcji logistycznej? Wyróżniamy 3 etapy:

- 1) na początku przez pewien czas jest bliska 0;
- 2) potem od pewnego momentu (wartości progowej) następuje nagły wzrost;
- 3) na końcu zbliża się do 1.

Model logistyczny zawdzięcza popularność następującym właściwościom:

1. Funkcja logistyczna przyjmuje wartości od 0 do 1. Model może opisywać wartości prawdopodobieństwa, które, jak wiemy, są zawsze zawarte między 0 a 1. Prawdopodobieństwa te określają najczęściej ryzyko zachorowania lub szanse wyzdrowienia. Dlatego model logistyczny jest często jedynym, jaki możemy wybrać, rozważając prawdopodobieństwa.
2. Kształt funkcji przypomina rozciągniętą literę S. Pokazuje on, że zmiany funkcji są minimalne, jeśli wartości zmiennych są mniejsze od pewnej wartości progowej. Gdy ją przekroczą, wówczas wartość funkcji zaczyna gwałtownie rosnąć do 1; prawdopodobieństwo utrzymuje się na wyjątkowo wysokim poziomie - blisko 1. Pojęcie wartości progowej jest często używane w badaniach medycznych i epidemiologicznych.

Zdefiniujemy teraz pojęcie "model logistyczny". Oznaczmy w tym celu przez Y zmienną dychotomiczną o wartościach:

- 1 - najczęściej dla zdarzeń pożądaných, na przykład: przeżycie, wyzdrowienie, niezakażenie, powodzenie leczenia;
- 0 - w przeciwnym wypadku, na przykład: zgon, choroba, zakażenie, niepowodzenie leczenia.

Wówczas model regresji logistycznej wiążący prawdopodobieństwo jednego z dwóch możliwych wyników zmiennej Y ze zmiennymi wyjaśniającymi jest określony równaniem

$$P(Y = 1 | x_1, x_2, \dots, x_k) = \frac{e^{(a_0 + \sum_{i=1}^k a_i x_i)}}{1 + e^{(a_0 + \sum_{i=1}^k a_i x_i)}}$$

gdzie

a_i $i = 0, \dots, k$ - to współczynniki regresji

$x_1, x_2, \dots, x_k$ - to zmienne niezależne, które mogą być mierzalne lub jakościowe.

Prawa strona równania to warunkowe prawdopodobieństwo, że zmienna Y przyjmie wartość 1 dla wartości zmiennych niezależnych $x_1, x_2, \dots, x_k$.

Jak w każdym modelu regresji, tak i tu staramy się oszacować współczynniki regresji $a_0, a_1, \dots, a_k$. Chcemy w ten sposób dopasować jak najlepszy model w oparciu o wartości pewnej grupy danych. Oczywiście liczebność grupy n musi być dostatecznie duża, co oznacza, że $n > 10(k + 1)$, gdzie k jest liczbą parametrów. W celu znalezienia estymatorów $a_0, a_1, \dots, a_k$ nie możemy stosować omówionej wcześniej metody najmniejszych kwadratów. Wymaga ona bowiem założenia o stałości wariancji, które w przypadku zmiennej dychotomicznej Y nie jest spełnione. W przypadku regresji logistycznej stosujemy metodę największej wiarygodności. Metoda ta, której podstawy teoretyczne opracował twórca analizy wariancji Fisher w 1929 roku, jest jedną z najważniejszych metod estymacji. Najogólniej mówiąc, metoda największej wiarygodności maksymalizuje funkcję wiarygodności, czyli iloczyn prawdopodobieństw pojawienia się poszczególnych obserwacji z próby przy danych parametrach modelu. Nie będziemy wchodzić w szczegółowe matematyczne rozważania; zwrócę tylko uwagę na istotę metody. Chodzi o to, aby jako ocenę szacowanych parametrów brać te wartości, dla których wiarygodność jest największa. Intuicyjne uzasadnienie takiego postępowania jest oczywiste - im większa wiarygodność (prawdopodobieństwo) zdarzenia, tym większa częstość względna (realizacja) w długim ciągu doświadczeń. W przypadku pojedynczego doświadczenia możemy i powinniśmy się liczyć z jego realizacją. Dla małej wiarygodności częstość względna jego wystąpienia jest bliska 0, i w przypadku pojedynczej próby możemy w ogóle nie brać pod uwagę możliwości jego realizacji. W przypadku konkretnego modelu im większa jest jego wiarygodność, tym większe prawdopodobieństwo, że wartości zmiennej zależnej pojawią się w próbie. Innymi słowy - im większa wiarygodność, tym lepsze dopasowanie modelu do danych. Estymatory wyznaczone metodą największej wiarygodności dzięki swoim własnościom gwarantują największe prawdopodobieństwo otrzymania zaobserwowanych wartości zmiennej zależnej. W statystycznych pakietach komputerowych estymatorów metody

największej wiarygodności poszukuje się, maksymalizując funkcję wiarygodności L lub jej logarytm. Z przyczyn obliczeniowych łatwiej jest bowiem znaleźć ekstremum funkcji $\log L$ niż samej funkcji L . Obliczenia dla konkretnego modelu mogą być skomplikowane (obliczamy bowiem prawdopodobieństwo pojawienia się każdej wartości y), jednak w tej żmudnej pracy wyręcza nas komputer. Używając pakietu STATISTICA, pokażemy, jak w praktyce przebiega taka estymacja i interpretacja otrzymanych wyników.

W pewnej klinice szukano czynników istotnie wpływających na przeżycie 90 dni po udarze mózgu. Zgodnie z wcześniejszymi badaniami i obserwacjami wzięto pod uwagę następujące zmienne:

WIEK - wiek pacjenta, zmienna ilościowa

SKP - stan kliniczny pacjenta przy przyjęciu, zmienna jakościowa (skala 0-58)

WIELCT - wielkość ogniska udaru, zmienna jakościowa (skala 1-4)

NAP - stężenie sodu w surowicy przy przyjęciu, zmienna ilościowa

CUKIERP - stężenie glukozy w surowicy przy przyjęciu, zmienna ilościowa

CPK - aktywność kinazy kreatynowej w surowicy przy przyjęciu, zmienna ilościowa

CHOL1 - stężenie cholesterolu w surowicy, zmienna ilościowa

P90 - przeżycie 90 dni, zmienna zależna dychotomiczna (1 - przeżycie, 2 - zgon).

Fragment arkusza danych dla 487 pacjentów przedstawia rysunek 2.

	1 P90	2 NAP	3 CUKIERP	4 CHOL	5 WIEL	6 WIEK	7 SKP	8 CPK
1	1	144	6.800	7.51	2	57	35	47.00
2	2	139	5.000	6.83	4	67	2	401.00
3	2	142	17.700	5.68	2	80	38	82.00
4	1	141	5.600	6.62	2	85	23	58.00
5	1	141	5.800	4.00	3	85	41	179.00
6	2	138	6.000	4.91	4	80		4159.00
7	1	141	4.900	6.86	3	48	54	57.00
8	1	135	6.400	5.35	4	78	28	165.00
9	1	143	7.400	4.57	1	41	30	
10	1	141	5.300	4.64	2	54	50	38.00
11	2	141	14.500	3.37	3	79	16	98.00
12	2	144	7.500	4.44	4	61	27	44.00

Rys. 2. Fragment arkusza danych

Ponieważ zmienna zależna P90 jest zmienną dychotomiczną, do analizy statystycznej użyjemy modelu regresji logistycznej.

W pakiecie STATISTICA analizę regresji logistycznej możemy przeprowadzić w module **Estymacja nieliniowa**. Po jego uruchomieniu otwiera się okno **Estymacja nieliniowa** przedstawione na rysunku 3.



Rys. 3. Okno wyboru modelu regresji

W oknie tym możemy wybrać pożądany typ regresji nieliniowej; interesuje nas regresja logistyczna. Po jej wybraniu i naciśnięciu klawisza **OK** pokazuje się okno **Regresja logistyczna** (rys. 4), otwierające procedurę analizy logistycznej. W oknie tym kolejno określamy:

- rodzaj danych na liście rozwijalnej **Plik wejściowy zawiera**;
- zmienne do analizy, za pomocą przycisku **Zmienne**;
- kody dychotomicznej zmiennej zależnej; zauważmy, że wszystkie wartości skojarzone z pierwszym kodem są równe zero (0), a więc jako pierwszy kod podajemy najczęściej symbol określający zdarzenie niepożądane, na przykład: zgon, choroba.

Regresja logistyczna (Logit)

Plik wejściowy zawiera: **Kody bez liczebności** [OK]

Zmienne [Anuluj]

Niezależna: NAP-CPK
Zależna: P90
Liczebności:

Kody zmiennej zależnej: **2,0** [Kliknij podwójnie wybrane pole, aby wybrać kody z listy wartości zmiennych.]
oraz **0,0**

Braki danych: **Usuwanie przypadkami** [SELECT CASES] [OK] [W]

Rys. 4. Okno wyboru zmiennych

Po wprowadzeniu i wybraniu zmiennych przyciskamy **OK**, aby otworzyć kolejne okno **Procedura estymacji** widoczne na rysunku 5.

Procedura estymacji

Model to: regresja logistyczna
Liczba estymowanych parametrów: 8
Funkcja straty: najw. wiarygodność
Zmienna zależna: P90
Kody: 2,0000 i 1,0000 (przek. odpowiednio na 0 i 1)
Zmiennie niezależne: NAP CUKIERP CHOL WIEL WIEK
Braki danych usuwano przypadkami
Liczba ważnych przyp.: 16

Metoda estymacji: **Rosenbrocka i quasi-Newtona** [OK] [Anuluj]

☒ Asymptotyczne błędy standardowe
☐ Eta dla aproks. skończ. różnic. 1.E- 8

Maksymalna liczba iteracji: 50
Kryterium zbieżności: .0001

Wartości początkowe: .1 dla wszystkich
Wstępna długość kroku: 1.0 dla wszystkich

Średnie i odchylenia standardowe
Wykres macierzowy ogółu zmiennych
Wykres ramkowy ogółu zmiennych

Rys. 5. Okno wyboru metody estymacji

W oknie tym możemy wybrać procedurę estymacji oraz określić kryterium zbieżności. Dokładne omówienie tych procedur i dostępnych tu opcji wychodzi poza ramy niniejszego opracowania. Zasygnalizujemy tylko dwa fakty. W celu obliczenia błędów standardowych estymowanych parametrów wybieramy opcję **Asymptotyczne błędy standardowe**.

Przed przeprowadzeniem właściwej analizy możemy po naciśnięciu przycisku **Średnie i odchylenia standardowe** przejrzeć podstawowe statystyki opisowe.

Po wprowadzeniu danych i wybraniu metody estymacji (tu Rosenbrocka i quasi-Newtona) arkusz wyników otrzymamy, naciskając przycisk **OK**, a następnie (w otwierającym się oknie **Wyniki**) przycisk **Parametry i błędy standardowe**. Pierwsze dwa arkusze wyników sygnalizowały, że zmienne NAP i CHOL1 okazały się nieistotne. A więc powinniśmy rozważyć prostszy model bez tych zmiennych. Otrzymane wartości przedstawia okno wyników (rys. 6). Ostatecznie nasz model przyjmuje postać jak we wzorze u dołu strony.

Model: Regresja logistyczna [logit] Liczba 0:124 1: 226						
ESTYM. NIELIN.	Znac. stat. P90 Strata: najwzrosty, błędów, skal. do 1 Koniec strata 139,07558616 Chi2(5)=176,89 p=0,0000 [3]					
N=350	Stosunek B0	CUKIERP	WIELCT	WIEK	SKP	CPK
Ocenia [1]	4,964	-,10214	-,34151	-,06256	,0812	-,00174
Błąd standard.	1,212	,03865	,16013	,01380	,0113	,00074
t(344)	4,095	-2,64248	-2,13277	-4,53159	7,1606	-2,34760
poziom p [2]	,000	,00861	,03365	,00001	,0000	,01946
-95% CL	2,580	-,17816	-,65646	-,08971	,0589	-,00320
+95% CL	7,348	-,02611	-,02656	-,03541	,1035	-,00028
Chi-kwadrat Walda	16,772	6,98270	4,54873	20,53528	51,2746	5,51123
poziom p	,000	,00823	,03295	,00001	,0000	,01890
Iloraz szans z jedn. [4]	143,155	,90291	,71070	,93936	1,0846	,99826
-95% CL	13,195	,83681	,51868	,91420	1,0607	,99681
+95% CL	1553,052	,97423	,97379	,96521	1,1091	,99972
Iloraz szans zakr.		,09352	,35896	,00204	111,2014	,00731
-95% CL		,01603	,13954	,00014	30,4848	,00012
+95% CL		,54563	,92340	,03004	405,6364	,45035

Rys. 6. Arkusz wyników estymacji parametrów modelu

Ujemne szacunki współczynników odpowiadające zmiennym CUKIERP, WIELCT, WIEK oraz CPK wskazują, że wzrost tych wielkości powoduje zmniejszenie się prawdopodobieństwa przeżycia 90 dni. Największy wzrost jest widoczny dla zmiennej WIELCT.

W arkuszu wyników znajdziemy również szereg dodatkowych statystyk i współczynników. Jak się nie zgubić w tym gąszczu liczb? Wskażę najważniejsze, na które przede wszystkim musimy zwrócić uwagę:

[1] współczynniki modelu regresji logistycznej umożliwiają zdefiniowanie modelu, jak w powyższym równaniu;

[2] wartości poziomu prawdopodobieństwa oceniającego istotność oszacowanych współczynników;

[3] wartość statystyki służąca do oceny dopasowania modelu - gdy statystyka χ^2 jest istotna (najczęściej gdy $p < 0,05$), to mówimy, że analizowany przez nas model wnosi coś nowego, różni się bowiem istotnie od modelu tylko z wyrazem wolnym;

[4] jednostkowy iloraz szans.

W modelu logistycznym dochodzi dodatkowy parametr - iloraz szans (*odds ratio*). Parametr ten jest często stosowany w badaniach klinicznych i epidemiologicznych. Przyjrzyjmy się więc mu bliżej. Przypomnijmy na początku pojęcie "szansy". Jest to stosunek prawdopodobieństwa, że jakieś zdarzenie wystąpi (np. rozwinię się rak płuca), do prawdopodobieństwa, że to zdarzenie nie wystąpi. Dla danego przypadku A powyższą definicję możemy zapisać następującym wzorem:

$$(\text{Szansa}) S(A) = \frac{p(A)}{p(\text{nie}A)} = \frac{p(A)}{1 - p(A)}$$

Na przykład, jeśli $p(A) = 0,2$ to $S(A) = 0,2/(1-0,2) = 1/4$. Oznacza to, że prawdopodobieństwo pojawienia się przypadku A jest równe jednej czwartej prawdopodobieństwa niepojawienia się tego przypadku. Możemy też powiedzieć, że szansa niewystąpienia przypadku A wynosi 4:1.

Iloraz szans dwóch grup porównywanych A i B definiujemy jako stosunek "szansy" wystąpienia A do "szansy" wystąpienia B, czyli $OR_{A \times B}$ (OR od *odds ratio*). Przy tak przyjętym oznaczeniu iloraz szans możemy zapisać w postaci następującego równania:

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

Na przykład oznaczmy jako A grupę palaczy papierosów, a jako B grupę osób niepalących. Przyjmując wartości $S(A) = 0,81$ i $S(B) = 0,2$, otrzymujemy iloraz szans "porównujący" szansę rozwinięcia się raka płuca u palących i niepalących: $OR_{A \times B} = 0,81/0,2 = 4,05$. Szansa rozwoju raka płuca u palaczy jest 4-krotnie większa niż u niepalących.

Zauważmy, że iloraz szans równy 1 oznacza równoważność ryzyka porównywanych grup. W powyższym przykładzie taka wartość mówiłaby, że palenie papierosów nie ma wpływu na rozwój raka płuca. Iloraz szans większy od 1 wskazuje, że szansa wystąpienia danego zdarzenia (np. zachorowania, zgonu itp.) w grupie A jest większa niż w grupie B. Odwrotnie - iloraz szans poniżej jedności wskazuje, że w grupie A szansa wystąpienia badanego

zdarzenia zdrowotnego jest mniejsza niż w grupie B.

Jednostkowy iloraz szans wyliczony w pakiecie pokazuje, jak zmienia się szansa danego zdarzenia przy jednostkowej zmianie zmiennej niezależnej.

W naszym przykładzie jednostkowy iloraz szans dla zmiennej WIELCT (wielkość ogniska udaru) jest równy 0,7107. W oparciu o tę wartość możemy zbudować tabelę obrazującą zmniejszanie się szansy przeżycia 90 dni wraz ze wzrostem wielkości ogniska udaru w skali od 1 do 4 (przy niezmienniających się pozostałych wartościach zmiennych). Ma ona postać następującą:

Tabela

wielkość ogniska	1	2	3	4
wielkość zmniejszenia		0,71	0,505	0,359

Widzimy, że chory, u którego wielkość ogniska wynosi 3, ma w stosunku do chorego z ogniskiem udaru wielkości 1 o połowę mniejszą szansę przeżycia 90 pierwszych dni od momentu zachorowania. Jeszcze mniejszą szansę (0,36) ma chory z największym ogniskiem udaru.

mgr Andrzej Stanisław

z Zakładu Biostatystyki i Informatyki Medycznej Collegium Medicum UJ w Krakowie

(Kierownik Zakładu: prof. dr hab. inż. Ryszard Tadeusiewicz)

Więcej informacji znajdą Państwo na stronie <http://www.mp.pl>

Copyright © 1996 - 2011 Medycyna Praktyczna