dr Izabela Miechowicz

Katedra i Zakład Informatyki i Statystyki

Zalecenia dotyczące regresji logistycznej

data przygotowania: 2024-05-31

Regresję logistyczną można zastosować w kilku celach:

- ✓ Kiedy chcesz sprawdzić, czy wyniki analiz jednowymiarowych pozostaną bez zmian w sytuacji, gdy porównywane grupy różnią się (nie są jednorodne) pod względem istotnych czynników takich jak np. wiek, płeć, masa ciała, itp.;
- ✓ Kiedy chcesz sprawdzić, jakie czynniki jednocześnie wpływają na wystąpienie analizowanego zjawiska;
- ✓ Kiedy chcesz zbudować model, na podstawie którego będziesz szacować prawdopodobieństwo.

Załóżmy, że badacz chce sprawdzić jakie zmienne spośród BMI w kategoriach wg WHO, glukozy, TSH oraz wartości zmodyfikowanej skali Ferrimana-Gallweya jednocześnie wpływają (i w jaki sposób) na wystąpienie choroby związanej z wystąpieniem zaburzeń hormonalnych u kobiet.

Co wiemy o danych?

(1) Zmienna **choroba** to zmienna dychotomiczna przyjmująca dwa stany: 0 – oznacza brak choroby, 1- oznacza wystąpienie choroby – jest to zmienna zmierzona na skali nominalnej

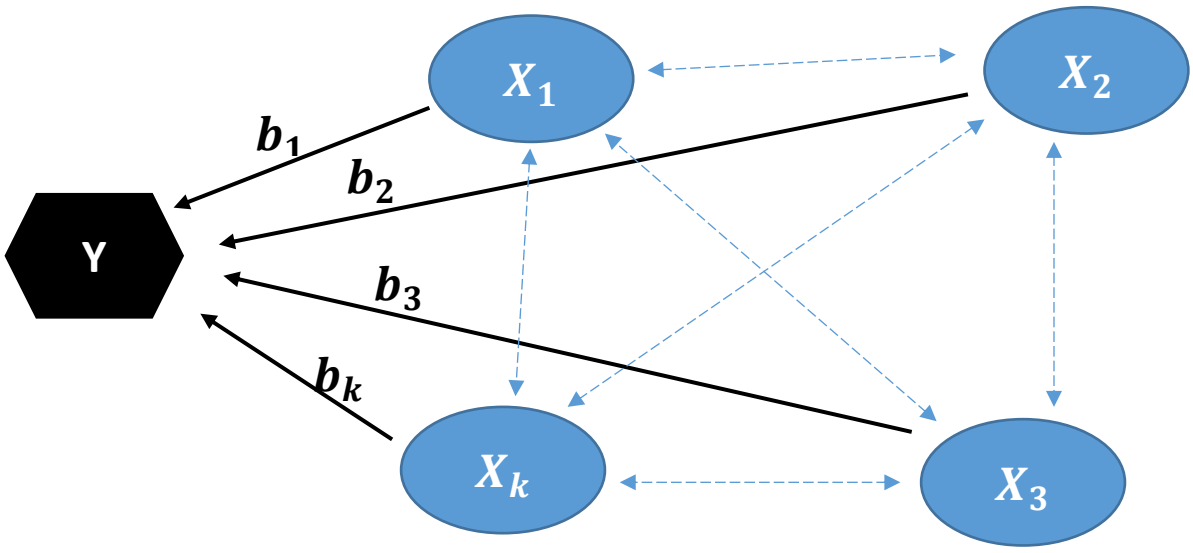
(2) **Kategorie BMI** można uporządkować, ale nie da się sprawdzić na ich podstawie o ile ktoś ma większy stosunek masy ciała do wzrostu, jedynie do której kategorii BMI należy – jest to skala porządkowa

(3) **Glukoza i TSH** to cechy dające się uporządkować i można obliczyć o ile jedna osoba ma wartości tych zmiennych wyższe od drugiej – jest to skala interwałowa a mierzony parametr ma charakter ciągły

(4) **Zmodyfikowana skala Ferrimana-Gallweya (mFG)** służy do oceny hirsutyzmu. Ocenia się owłosienie 9 okolic ciała, każdą w skali od 0-4pkt. Następnie punkty należy zsumować – jest to skala porządkowa

Założenia, które musimy sprawdzić

(1) Założenia dotyczące danych:



Zmienne niezależne powinny być skorelowane ze zmienną niezależną, a jednocześnie mogą być słabo skorelowane z sobą.

(2) Założenia dotyczące liczności:

n>10 (k + 1),
gdzie n-liczebność grupy, k-oznacza liczbę parametrów włączonych do modelu

lub

n>=10k/p,
gdzie p oznacza mniejszą z proporcji liczności opisanej ze zmiennej zależnej (tzn. proporcji zdrowych lub chorych)

Jak przygotować dane?

Dane przygotuj w kolumnach

choroba	BMI kategorie	Glukoza mg/dL	mFG (0-36 pkt)	TSH uIU/mL
1	4	80	15	2,42
0	1	75	9	4,24
0	3	84	10	1,65
1	2	73	7	3,54
…	…	…	…	…

Dodatkowo, kategorie BMI można rozbić na tzw. zmienne fikcyjne (dummy variables), przyjmując normę (kategoria 2) za kategorię referencyjną. Wówczas wyniki analizy regresji będą interpretowane zawsze w stosunku do kategorii referencyjnej. W takim przypadku dane będą wyglądały następująco:

choroba	BMI kategorie	niedowaga	nadwaga	Glukoza mg/dL	mFG (0-36 pkt)	TSH uIU/mL
1	2	0	0	80	15	2,42
0	1	1	0	75	9	4,24
0	3	0	1	84	10	1,65
1	2	0	0	73	7	3,54
…	…			…	…	…

Jak interpretować wyniki?

Pamiętaj, że **wartość α** najczęściej wynosi 0.05 i jest to poziom błędu którego nie chcemy przekroczyć prowadząc badanie.

Istotność modelu – oceniamy testem ilorazu wiarygodności, który uwzględnia istotność całego modelu, a nie pojedynczych składowych i daje odpowiedź na pytanie, czy model zawierający zmienną (bądź zmienne) niezależną pozwala na lepsze przewidywanie wyników w porównaniu z modelem, który tej zmiennej nie zawiera. Zwykle modelem odniesienia jest model zerowy, tzn. zawierający tylko wyraz wolny.

Istotność zmiennych budujących model – oceniana za pomocą testu chi^2 Walda, który pozwala określić, które ze zmiennych włączonych do modelu i jak silnie oddziałują na zmienną zależną.

Iloraz Szans (OR-Odds Ratio) - pozwala określić jaki wpływ na zmienną zależną ma dana zmienna niezależna - INTERPRETACJA

- ✓ jeśli OR>1, to wzrost danej zmiennej o jednostkę zwiększa szansę na wystąpienie danego zjawiska;
- ✓ jeśli OR<1, to wzrost danej zmiennej o jednostkę zmniejsza szansę na wystąpienie danego zjawiska.

Uwaga!

Zanim będziesz interpretować iloraz szans upewnij się, że wyznaczony dla niego 95% przedział ufności (95% CI – confidence interval) nie zawiera 1. Przedział ufności informuje, że z prawdopodobieństwem 95% OR w populacji będzie znajdował się w tym przedziale.

Łatwiej o interpretację, gdy OR>1. Zatem w sytuacji gdy $OR_{AxB} < 1$, możesz dokonać przekształcenia w następujący sposób:

$$OR_{BxA} = 1 / OR_{AxB}$$

W analogiczny sposób należy również przekształcić 95% ufności, tzn. gdy przedział ufności dla OR_{AxB} wynosi (a; b) to dla OR_{BxA} będzie wynosił $(\frac{1}{b}; \frac{1}{a})$.

Jak ocenić jakość zbudowanego modelu?

- ✓ Można interpretować **diagnostyczny iloraz szans**. Pozwala on określić ile razy więcej przypadków zostało zaklasyfikowanych poprawnie niż błędnie, przy pomocy zbudowanego modelu. Obliczysz go dzieląc przez siebie iloczyn poprawnie zaklasyfikowanych zer i jedynek przez iloczyn błednie zaklasyfikowanych zer i jedynek.

- ✓ **Test Hosmera-Lemeshowa -** zaproponowany test dobroci dopasowania, oparty na teście chi^2, porównuje wartości rzeczywiste z wartościami oczekiwanymi wyznaczonymi z modelu. Dlatego też pożądane jest uzyskanie braku istotności współczynnika chi^2.

- ✓ Współczynniki (pseudo R^2, Nagelkerke, Coxa-Snella) – im wyższe wartości współczynników, tym lepsze przewidywanie przez model zmiennej zależnej.

- ✓ Ocena klasyfikacji dokonanej przez model: oceniamy odsetek poprawnie zaklasyfikowanych chorych (czułość), procent poprawnie zaklasyfikowanych osób zdrowych (swoistość) oraz ogólny procent poprawnie zaklasyfikowanych przypadków.

- ✓ Pole pod krzywą ROC (AUC - area under curve) – sprawdzamy czy jego wartość jest istotnie większe od 0,5

Jak opisać wyniki?

- ✓ W przypadku poszczególnych zmiennych podaj wartość p, OR oraz 95% przedział ufności
- ✓ Podaj wartość p w teście ilorazu wiarygodności, określ jakość modelu.
- ✓ Określ jakość klasyfikacyjną modelu podając jego czułość i swoistość wraz z 95% przedziałami ufności.

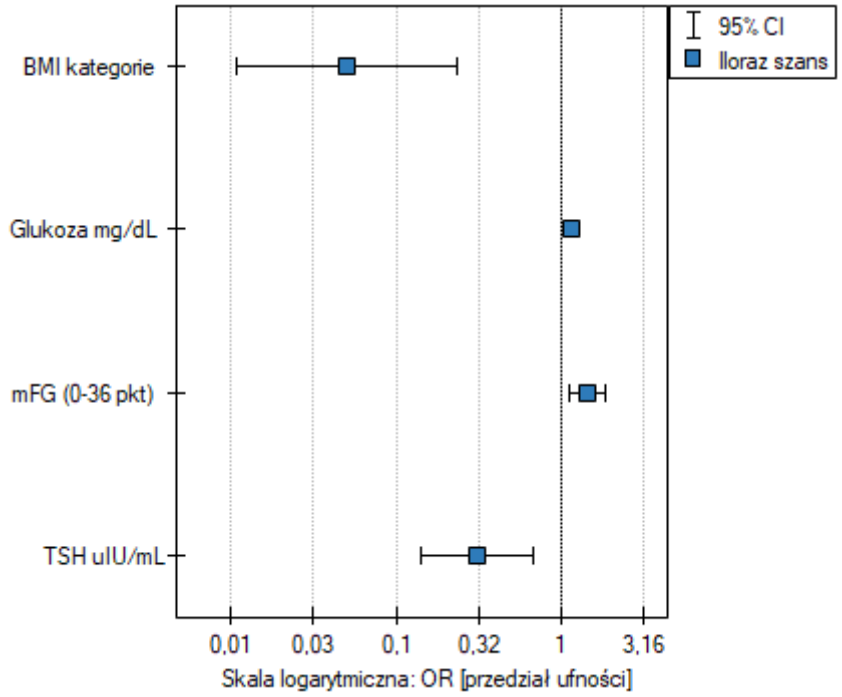
Przykładowy wynik

zmienne	wartość p	iloraz szans	-95% CI	+95% CI
w. wolny	0,065	0,001	<0,001	1,492
bmi kategorie	<0,001	0,050	0,011	0,232
glukoza mg/dl	0,001	1,148	1,058	1,246
mFG (0-36 pkt)	0,004	1,442	1,125	1,847
tsh uIU/mL	0,004	0,310	0,140	0,685

Wszystkie zmienne budujące model istotnie wpływają na wystąpienie choroby. W przypadku BMI w kategoriach oraz TSH OR<1 (w oby przypadkach 95% przedział ufności nie zawiera 1), zatem wzrost tych zmiennych o jednostkę zmniejsza szansę na wystąpienie choroby (to są destymulanty). W przypadku glukozy i mFG OR>1 (również przedziały ufności nie zawierają 1), zatem wzrost

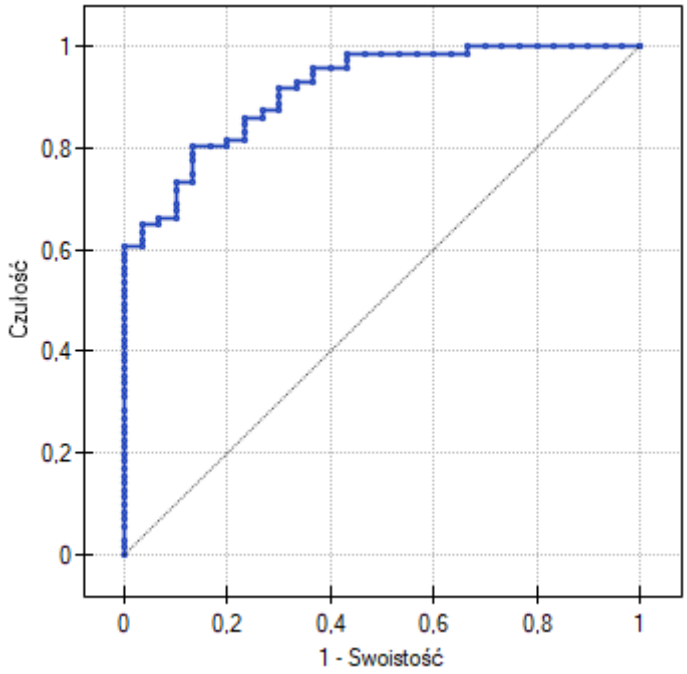
tych zmiennych o 1 zwiększa szanse na wystąpienie choroby (w przypadku glukozy 1,15 razy, a dla mFG 1,44 razy) – są to stymulanty.

Bardzo przydatny do powyższej analizy może być wykres przedstawiający ilorazy szans wraz z przedziałami ufności dla każdej zmiennej. Można z niego wyczytać nie tylko to, czy przedział ufności zawiera 1, ale również zaobserwować ich szerokość.



Jakość modelu	
Test ilorazu wiarygodności p	<0,001
Pseudo R2	0,467
R2(Nagelkerke)	0,616
R2(Coxa-Snella)	0,433
Hosmer-Lemeshow test p	0,572
Krzywe ROC (DeLong's method)	
AUC	0,915
-95% CI	0,862
+95% CI	0,969
Wartość p	<0,001

Test ilorazu wiarygodności wskazał, że zbudowany model jest istotny statystycznie (p<0,001). Test Hosmera-Lemeshowa wskazuje, że nie ma różnic pomiędzy wartościami rzeczywistymi a wartościami oczekiwanymi wyznaczonymi z modelu (p=0,572). Dodatkowo uzyskano wysoką (0,915) i istotnie różną od 0,5 (p<0,001) wartość pola pod krzywą ROC.



	WARTOŚCI OBSERWOWANE	
WARTOŚCI PRZEWIDYWANE	**1**	**0**
1	65	9
0	6	21

Diagnostyczny iloraz szans=(65*21)/(6*9)=1365/54=25, czyli 25 razy więcej przypadków zostało zaklasyfikowanych poprawnie niż błędnie.

Klasyfikacja

% poprawnie zaklasyfikowanych przypadków	85%
Czułość (% poprawnych 1)	92%
-95% CI	83%
+95% CI	97%
Swoistość (% poprawnych 0)	70%

-95% CI	51%
+95% CI	85%

Spośród wszystkich 101 przypadków model zaklasyfikował poprawnie 85% ((65+21)/101). Czułość modelu wyniosła 92% (95%CI=(83%, 97%), co oznacza, że spośród wszystkich chorych model rozpoznał poprawnie 92% (65/71). Swoistość modelu wyniosła 70% (95%CI=(51%, 85%)), czyli spośród wszystkich osób zdrowych model rozpoznał poprawnie 70% (21/30).

W przypadku, gdybyśmy chcieli analizować wpływ BMI na wystąpienie choroby, ale w odniesieniu do wartości referencyjnej jaką jest norma, wyniki wyglądałyby następująco:

	wartość p	iloraz szans	-95% CI	+95% ci
w. wolny	0,001	<0,001	<0,001	0,007
niedowaga	0,001	27,498	4,211	179,550
nadwaga	0,102	0,109	0,008	1,555
glukoza mg/dl	0,001	1,149	1,059	1,247
mFG (0-36 pkt)	0,003	1,449	1,132	1,855
tsh uIU/mL	0,004	0,299	0,133	0,676

Teraz, analizując wpływ BMI na wystąpienie choroby można zauważyć, że niedowaga ponad 27 razy zwiększa szansę na jej wystąpienie w porównaniu do normy, natomiast nadwaga nie ma wpływu na wystąpienie choroby. Charakter pozostałych zmiennych nie uległ zmianie.