Iwona Mróz,

Instytut Fizyki Doświadczalnej,

Uniwersytet Wrocławski

**WARSZTATY – REGRESJA LOGISTYCZNA**

Niniejsze materiały mają charakter roboczy. Bardzo proszę o zgłaszanie zauważonych błędów, braków, niedociągnięć i niejasności. Prośba dotyczy też przypisów. Z góry dziękuję za pomoc☺.

*Tekst opracowano głównie na podstawie podręcznika Andrzeja Stanisza „Przystępny kurs statystyki z zastosowaniami STATISTICA.PL na przykładach z medycyny”, tom 2 „Modele liniowe i nieliniowe”, StatSoft, Kraków 2007, str. 217-226. Pochodzą z niego informacje, do których nie podano innych źródeł. Opisany w punkcie 2 przykład jest wzorowany na przykładzie ze str. 222-224.*

**1. Metoda największej wiarygodności**

Najpopularniejszą metodą szacowania parametrów modeli regresyjnych jest metoda najmniejszych kwadratów (MNK). Nie może ona być jednak stosowana w każdym przypadku, np. ze względu na brak możliwości spełnienia założeń MNK tak jak ma to miejsce w przypadku opisanej w punkcie 2 regresji logistycznej. Inną metodą pozwalającą na estymację parametrów modelu regresyjnego jest metoda największej wiarygodności (*ang. maximum likelihood*) na szukaniu maksimum funkcji wiarygodności L określonej jako:

,

gdzie określa prawdopodobieństwo pojawienia się wartości zmiennej zależnej równej *yi*przy zestawie parametrów .[[1]](#footnote-1)

Przystępny przykład zastosowania metody największej wiarygodności do znalezienia parametru λ rozkładu Poissona przedstawiono na filmie dostępnym pod adresem:

<https://www.youtube.com/watch?v=1BXMgzas4g0>.[[2]](#footnote-2) Poniższe rozważania przytaczam za Panią prowadzącą wykład zarejestrowany na filmie.

Niech zmienna losowa X podlega rozkładowi Poissona, czyli:

Dysponujemy *n*-elementową próbą prostą , której realizacją jest .

Funkcja największej wiarygodności ma postać:

Naszym zadaniem jest znalezienie takiej wartości parametru , dla której funkcja *L* osiąga maksimum. Ze względów obliczeniowych, zamiast maksymalizować funkcję *L*, maksymalizujemy wyrażenie *lnL*:

Wyrażenie *lnL* osiąga maksimum jeśli i . Obliczamy:

Dla wyliczonego powyżej parametru funkcja *lnL* osiąga maksimum, gdyż:

Dla danej realizacji próby prostej wartość jest estymatorem parametru rozkładu Poissona.

**2. Regresja logistyczna**

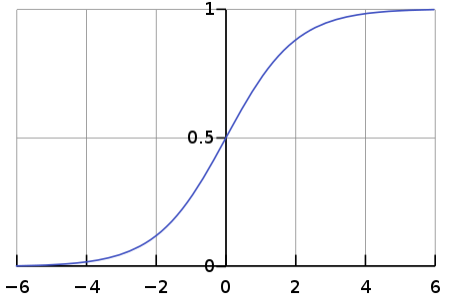
W modelach regresyjnych, które do tej pory rozważaliśmy, badana cecha statystyczna opisywana przez zmienną zależną miała charakter ilościowy i ciągły. Predyktory (zmienne niezależne) mogły być ilościowe lub jakościowe (sposób wprowadzania zmiennych jakościowych do modelu regresyjnego przez ich odpowiednie kodowanie jeszcze omówimy). Obecnie rozważymy model regresyjny, w którym zmienna objaśniana ma charakter dychotomiczny i przyjmiemy, że może ona przyjmować tylko dwie wartości: 0 i 1.

Zanim przejdziemy do formalnego opisu modelu regresji logistycznej, przyjrzyjmy się krzywej logistycznej, której równanie stanowi podstawę do budowy tego modelu.

Równanie tzw. standardowej krzywej logistycznej ma postać:[[3]](#footnote-3)

{\displaystyle Y\_{i}\ \sim B(p\_{i},n\_{i}),} (\*)

A jej wykres wygląda następująco:



Źródło: Wikipedia, <https://pl.qaz.wiki/wiki/Logistic_function>, data dostępu: 23.04.2021.

Zauważmy, że dziedziną funkcji jest zbiór liczb rzeczywistych, a zbiór wartości przebiega od 0 do 1, przy czym przejście od wartości bliskich zeru do wartości bliskich 1 zachodzi dla stosunkowo wąskiego przedziału dziedziny, którego środkiem jest *x* = 0.

Charakterystyczny, przypominający literę S kształt krzywej logistycznej może być modyfikowany przez wprowadzenie do równania parametrów, takich jak wartość punktu środkowego krzywej sigmoidalnej (*x0*), maksymalnej wartości osiąganej przez funkcję (*L*) i tzw. logistycznego tempa wzrostu opisującego „stromość” krzywej (*k*). W przypadku standardowej krzywej logistycznej *x0* = 0, *L* = 1 i *k* = 1. W innym przypadku równanie (\*) przybiera postać:

(\*\*)

Fakt, że zbiorem wartości krzywej logistycznej może być przedział od 0 do 1 oraz możliwość kontrolowania „stromości” krzywej przy przechodzeniu od wartości bliskich 0 do wartości bliskich 1 sprzyja konstrukcji modelu, w którym dany zestaw wartości wielu predyktorów z pewnym prawdopodobieństwem prowadzi do sukcesu (zmienna objaśniana przybiera wartość 1) lub porażki (zmienna objaśniana przybiera wartość zero). Model taki może zostać wykorzystany do analizowania zagadnień o charakterze klasyfikacyjnym i do oceny prawdopodobieństwa, czy dla danego zestawu wartości predyktorów zajdzie zdarzenie traktowane jako sukces.

Określmy zestaw k zmiennych losowych *X1*, *X2*, …, *Xk*, które przybierają wartości *x1*, x2, …, *xk*, i które pozwalają na wyznaczenie prawdopodobieństwa sukcesu *P*(*Y*=1). Zestaw zmiennych losowych traktujemy jak wektor losowy ***X***.

Prawdopodobieństwo wystąpienia sukcesu określamy jako:

gdzie *a0, a1, a2,…,ak* to współczynniki regresji, które chcemy wyznaczyć budując model.

Szczegółowe rozważania wskazują, że liczebność próby *n* wykorzystanej do zbudowania modelu regresji logistycznej musi być duża. Przyjmuje się, że , gdzie *k* jest liczbą parametrów[[4]](#footnote-4). Ponadto, dla zmiennej dychotomicznej o wartości oczekiwanej μ wariancja nie jest stała i wynosi Nie będzie zatem spełnione założenie o stałości wariancji, co nie pozwala na wykorzystanie metody najmniejszych kwadratów do wyznaczenia parametrów modelu. Dlatego w regresji logistycznej parametry wyznacza się metodą największej wiarygodności.

Użyteczną postacią modelu regresji logistycznej jest jej postać logitowa powstała wskutek transformacji:

[[5]](#footnote-6),

gdzie po prawej stronie równania widzimy zależność liniową.

Postać logitowa modelu logistycznego pozwala na łatwą interpretację bardzo ważnego parametru, który nazywamy ilorazem szans (*ang. odds ratio*). Iloraz szans pozwala na porównanie szansy zajścia danego zjawiska w porównywanych grupach.

Szansą S(A) wystąpienia zdarzenia A nazywamy:

Zauważmy, że prawdopodobieństwo i szansa to różne pojęcia, szansa jest ilorazem prawdopodobieństw.

Iloraz szans pozwala porównać szanse wystąpienia zdarzenia A w dwóch porównywanych grupach. Niech B1 i B2 oznaczają odpowiednio wystąpienie zdarzenia w grupach 1 i 2. Obliczymy go jako:

Jeżeli wartość > 0 to szansa wystąpienia zdarzenia A w grupie B1 jest większa niż w grupie B2, jeżeli to szanse wystąpienia A w obu grupach są takie same, a jeżeli to szansa wystąpienia A w grupie B1 jest mniejsza niż w grupie B2.

Iloraz szans to zmienna losowa. Jej wartość obliczona na podstawie danych z próby jest estymatorem wartości ilorazu szans w populacji.

Przykład (dane są fikcyjne, wnioski mogą być sprzeczne z rzeczywistymi ustaleniami):

Podczas pandemii spowodowanej SARS-COV-2 sprawdzano, czy płeć dorosłych osób wpływa na częstość zachorowań na COVID-19. W tym celu obserwowano przez określony czas 1500 dorosłych osób: 865 kobiet i 635 mężczyzn. Podczas eksperymentu zachorowało 187 kobiet i 154 mężczyzn. Trzeba określić, czy i ile razy szansa zachorowania na COVID-19 jest większa (mniejsza) w przypadku kobiet.

Przedstawmy dane w postaci tabeli:

|  |  |  |
| --- | --- | --- |
| Płeć | COVID-19 | |
| Zachorowało | Pozostało zdrowych |
| Kobiety | 187 | 865-187=678 |
| Mężczyźni | 154 | 635-154=481 |

Niech rozpatrywanym zdarzeniem będzie zachorowanie na COVID-19. Oznaczmy grupę kobiet jako B1, a grupę mężczyzn jako B2. Szansa na zachorowanie w grupie kobiet wynosi: S(B1) = (187/865)/(678/865) = 0,276, a w grupie mężczyzn: S(B2) = (154/635)/(481/635) = 154/635 = 0,243. Stąd, iloraz szans: = 0,276/0,243 = 1,136. Zatem, kobiety chorują 1,136 razy częściej niż mężczyźni.

Przyjrzyjmy się teraz związkowi ilorazu szans z modelem regresji logistycznej. Jeżeli za zdarzenie A przyjmiemy Y=1, to możemy zapisać:

Z drugiej strony wiemy, że:

,

a zatem (\*\*\*).

Jeżeli chcemy obliczyć iloraz szans dla wystąpienia zdarzenia w grupach 1 i 2 (B1 i B2), do wzoru (\*\*\*) podstawiamy wartości *xi* dla poszczególnych grup i wyliczamy szanse dla grup, a następnie obliczamy .

1. A. Stanisz, s. 220-221. [↑](#footnote-ref-1)
2. Strona internetowa „Matematyka na plus”, <https://www.youtube.com/watch?v=1BXMgzas4g0>, data ostatniego dostępu: 22.05.2022. [↑](#footnote-ref-2)
3. A. Stanisz, s. 218. [↑](#footnote-ref-3)
4. Ibidem, s. 220. [↑](#footnote-ref-4)
5. Ibidem, s. 223. [↑](#footnote-ref-6)