

PETRESCU Mary-Violeta

INTERPRETAREA PROBABILITĂȚII LOGARITMICE ȘI A RAPORTULUI DE PROBABILITATE

Craiova
Septembrie, 2025

Copyright © 2025, Autor
Tehnoredactare computerizată: PETRESCU Mary-Violeta
Corectura aparține autorului

Responsabilitatea pentru conținutul lucrării revine în exclusivitate autorului
Studiu științific

Publicat: Septembrie 2025
<http://www.edituraelse.ro/ped/ped.htm>

INTERPRETAREA PROBABILITĂȚII LOGARITMICE ȘI A RAPORTULUI DE PROBABILITATE

Coeficienții logit standardizați sunt coeficienții „b” (valorile pantei) din ecuația de regresie. Panta poate fi interpretată drept schimbarea în valoarea medie a lui Y, dintr-o unitate de schimb în X.

Regresia logistică calculează schimbările în probabilitatea logaritmică a variabilei dependente, nu modificările în valoarea variabilei dependente, așa cum face regresia celor mai mici pătrate (OLS).

Pentru o variabilă dihotomică, șansele de a fi membru al grupului țintă sunt egale cu probabilitatea de a fi membru în grupul țintă împărțit la probabilitatea de a fi membru în grupul celălalt.

Un concept important îl reprezintă raportul de probabilitate (OR), care estimează schimbările în șansele apartenenței la grupul țintă pentru o creștere de o unitate la predictor, fiind calculat prin folosirea unui coeficient de regresie al predicției ca fiind exponentul sau exp.

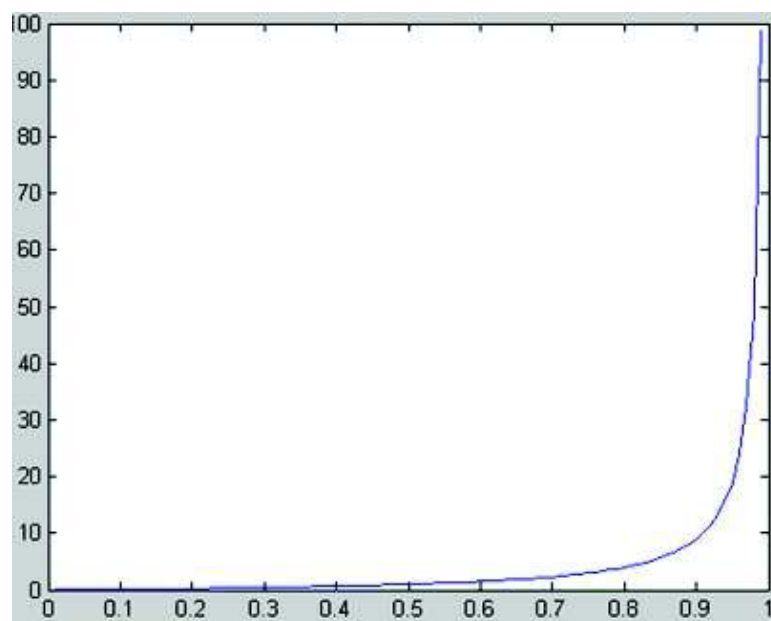


Figura 1. Valoarea raportului de probabilitate

Valoarea raportului de probabilitate (a se vedea Figura 1.) poate varia de la 0 la ∞ și indică cât de probabil este ca o observație să fie mai degrabă parte din grupul țintă decât membru al celuilalt grup.

La fel ca în cazul regresiei liniare, încercăm să găsim o linie de potrivire perfectă, dar pentru că valorile lui Y pot varia doar între 0 și 1, nu putem folosi metoda celor mai mici pătrate.

Verosimilitatea maximă (ML) este utilizată pentru a afla funcția care va maximiza abilitatea noastră de a predicta probabilitatea lui Y bazată pe ceea ce știm despre X. Cu alte cuvinte, (OR) găsește cele mai bune valori pentru formula (3.2.2).

Probabilitatea se referă la șansele care există. Întotdeauna înseamnă probabilitate sub o ipoteză specifică.

În regresia logistică, sunt importante două ipoteze:

- ipoteza nulă, atunci când toți coeficienții din ecuația de regresie iau valoarea zero;
- ipoteza alternativă care arată că modelul cu predictorii ce este luat acum în considerare este precis, și diferă semnificativ de mulțimea vidă z , de exemplu oferă în mod semnificativ ceva mai mult decât probabilitatea și nivelul de predicție întâmplător al ipotezei nule.

Apoi, lucrăm la probabilitatea observării datelor pe care chiar le-am observat sub fiecare din aceste ipoteze. Rezultatul este de obicei un număr foarte mic, și pentru a fi mai accesibil, este folosit logaritmul natural, producând o verosimilitate logaritmică (LL).

Probabilitățile sunt întotdeauna mai puțin de una, așa că LL este întotdeauna negativă. Acestea reprezintă baza pentru testele unui model logistic.

Testul raportului de probabilitate este bazat pe raportul $-2LL$. Este un test al semnificației diferenței dintre raportul de probabilitate $-2LL$ pentru modelul de cercetare cu predictorii, numit modelul chi pătrat (Chi-Square sau χ^2) minus raportul de probabilitate pentru modelul standard cu doar o constantă în el.

Semnificația la nivelul .05 sau mai scăzut înseamnă că modelul de cercetare cu predictorii este diferit în mod semnificativ față de cel cu doar o constantă (toți coeficienții “b” fiind zero). Aceasta măsoară îmbunătățirea pe care o fac variabilele explicative în comparație cu modelul nul, iar χ^2 este folosit pentru a conferi semnificație acestui raport.

Când probabilitatea eşuează să ajungă la nivelul de semnificație de 5%, păstrăm ipoteza nulă, știind că variabilele independente (predictorul) nu au efecte de creștere (nu fac nicio diferență) în a predica variabila dependentă.

În practică, putem simplifica o notație având doar un singur parametru β în loc de doi parametri α și β .

Dacă problema noastră inițială este formulată astfel încât: $y = \alpha + x \cdot \beta + \varepsilon$, vom rescrie aceasta ca: $y = [1, x] \cdot [\alpha, \beta]^T + \varepsilon$.

Dacă notăm acum $\beta' = [\alpha, \beta]^T$ și $x' = [1, x]$, atunci vom putea să formulăm aceeași problemă cu numai „un” parametru de model β' : $y = x' \cdot \beta' + \varepsilon$.

Dar aceasta nu este altceva decât o schimbare de notație.

Avem nevoie de încă doi parametri pentru a determina, dar simplificând notația noastră, astfel încât să avem nevoie acum pentru a estima parametrul β' .

De acum, vom nota β' ca β și x' ca x și problema noastră va fi, prin urmare, de a obține parametrul modelului, β când $y = x \cdot \beta + \varepsilon$.

Dacă am făcut n observații, cu y_i răspunsuri și x_i predictorii, putem defini:

$$Y = [y_1, y_2, \dots, y_n]^T$$

$$X = [x_1, x_2, \dots, x_n]^T.$$

Sistemul pe care dorim să-l rezolvăm pentru a găsi parametrul β este scris ca:

$$Y = X \cdot \beta$$

Soluția minimă a erorii pătratice a acestui sistem se află după cum urmează:

$$Y = X \cdot \beta$$

$$X^T \cdot Y = X^T \cdot X \cdot \beta$$

$$\beta = (X^T \cdot X)^{-1} \cdot X^T \cdot Y.$$

Avem nevoie doar pentru a evalua expresia $(X^T \cdot X)^{-1} \cdot X^T \cdot Y$ și am găsit β care minimizează suma pătratelor rezidurilor.

Regresia logistică este în general considerată ca o metodă de modelare în situații pentru care există o variabilă de răspuns binar.

Variabilele predictor pot fi numerice sau categoriale (inclusiv binar).

Regresia logistică multiplă poate fi utilizată atunci când există mai mult de două rezultate posibile pentru răspuns.

În continuare ne vom axa doar pe răspunsul binar.

Considerăm Y variabila de răspuns binar și presupunem că $P(Y = 1)$ este posibil dependent de $\vec{x}$, un vector cu valorile predictorului.

Scopul este de a modela:

$$p(\vec{x}) \equiv P(Y = 1 | \vec{x}) \quad (1)$$

Deoarece Y este binar, modelarea lui $p(\vec{x})$ este realmente modelarea lui $E(Y | \vec{x})$, care este exact ceea ce face metoda celor mai mici pătrate cu răspuns numeric.

Metoda celor mai mici pătrate este o metodă folosită pentru estimarea parametrilor necunoscuți în modelul regresiei liniare.

Dacă avem un model $p(\vec{x})$ ca o funcție liniară cu variabilele predictor (variabile de regresie) $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$, atunci modelul poate duce la probabilități estimate care sunt în afara intervalului $[0,1]$. Pentru a funcționa mai bine se presupune că:

$$p(\vec{x}) = \frac{\exp^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + \exp^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \quad (2)$$

unde $x_1, \dots, x_p$ poate fi setul original de variabile explicative, dar predictorii pot să includă variabile transformate și construite.

De exemplu, în cazul în care există doar două variabile explicative, și, apoi un model de ordinul întâi care tocmai folosește v_1 și v_2 ca predictorii poate să nu fie adecvat, dar un model de ordinul al doilea folosind $x_1 = v_1$, $x_2 = v_2$, $x_3 = v_1^2$, $x_4 = v_2^2$, & $x_5 = v_1 v_2$ poate oferi o potrivire bună.

Se poate observa că:

$$\log\left(\frac{p(\bar{x})}{1-p(\bar{x})}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (3)$$

unde $\log(p(\bar{x})/[1-p(\bar{x})])$ este numit logit.

În timp ce modelul pentru $p(\bar{x})$ este ceva mai complicat, modelul pentru logit este liniar la predictorii.

De asemenea, indiferent de ce valoare este $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$ estimarea corespunzătoare a $p(\bar{x})$ va fi cuprinsă între 0 și 1.

Parametrii necunoscuți (coeficienții $\beta_0, \beta_1, \dots, \beta_p$) sunt în general estimați de maximizarea probabilității, $\prod_{i=1}^n \{p(\bar{x}_i)^{y_i} [1-p(\bar{x}_i)]^{1-y_i}\}$, care este doar o expresie pentru $P(Y_1 = y_1, \dots, Y_n = y_n | \bar{x}_1, \dots, \bar{x}_n)$.

BIBLIOGRAFIE:

1. Gunn, S.R., 1998. Support Vector Machines for Classification and Regression. University of Southampton, 1998.
2. Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer - Verlag, ISBN 0-387-98780-0.