PETRESCU Mary-Violeta

CONSTRUIREA UNUI MODEL CU UNA SAU MAI MULTE VARIABILE DE RĂSPUNS BINAR

Craiova
Februarie, 2025

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

Responsabilitatea pentru conţinutul lucrării revine în exclusivitate autorului
Studiu științific

Publicat: februarie 2025
http://www.edituraelse.ro/ped/ped.htm

CONSTRUIREA UNUI MODEL CU UNA SAU MAI MULTE VARIABILE DE RĂSPUNS BINAR

În lumea afacerilor există o serie de variabile interesante care sunt dihotomice, cum ar fi : consumatorii iau decizii dacă să cumpere sau nu, un produs poate trece cu bine sau poate să eşueze într-un control al calităţii, există riscuri de credit bune sau rele. Regresia logistică este de regulă preferată în locul analizei discriminante în cazul în care există doar două categorii de variabile dependente.

Regresia logistică utilizează teoria binomială a probabilităţii, în care sunt doar două valori ce se pot prezice, adică probabilitatea (p) este 1 şi nu 0, adică evenimentul/persoana aparţine numai unui grup şi nu celuilalt. Regresia logistică formează cea mai bună ecuaţie sau funcţie folosind metoda verosimilităţii maxime, care maximizează probabilitatea clasificării datelor observate în categorii potrivite, dată fiind regresia coeficienţilor.

Există două mari utilizări ale regresiei logistice:

- prima este anticiparea membrilor grupului deoarece regresia logistică calculează probabilitatea de succes raportată la probabilitatea de eşec, rezultatele analizei sunt sub forma unui raport de probabilitate (oddsratio);
- regresia logistică oferă, de asemenea, informaţii despre relaţiileşiforţele dintre variabile.

Premisele regresiei logistice sunt:

- regresia logistică nu porneşte de la premisa că există o relaţie liniară între variabilele dependente şi cele independente;
- variabilele dependente trebuie să fie o dihotomie (două categorii);
- variabilele independente nu trebuie să fie în mod necesar nici intervale, nici distribuite normal, nici liniar legate, şi nici dispersate în mod egal în cadrul fiecărui grup;
- categoriile (grupurile) trebuie să se excludă reciproc şi complete, un caz poate exista doar într-un grup şi fiecare caz trebuie să fie un membru al unuia din grupuri.

- sunt necesare eşantioane de date mai mari decât ca pentru regresia liniară deoarece coeficienţii maximi de probabilitate sunt estimaţi la un număr mare de modele (este recomandat un număr minim de 50 de cazuri pe predictor).

Ecuaţia regresiei logistice

În timp ce regresia logistică dă fiecărui dispozitiv de predicţie un coeficient "b" care măsoară contribuţia sa independentă la variaţii în variabila dependentă, variabila dependentă poate lua doar una din cele două valori, adică 0 sau 1.

Ceea ce vrem să anticipăm din informaţia variabilelor relevante independente şi a coeficienţilor este prin urmare, nu o valoare numerică a unei variabile dependente ca în regresia liniară, ci mai degrabă probabilitatea (p), care este mai degrabă 1 decât 0 (aparţinând numai unui grup şi nu, celuilalt).

Dar chiar şi folosirea probabilităţii ca şi variabilă dependentă este defectuoasă, în principiu pentru că predictorii numerici pot apărea într-o gamă nelimitată.

Daca l-am exprima pe (p) drept o funcţie liniară de investiţie, e posibil să ne aflăm în situaţia în care predictăm că p este mai mare decât 1 (ceea ce nu poate fi adevărat, deoarece probabilităţile pot lua valori doar între 0 şi 1).

În plus, pentru că regresia logistică are doar două valori y – în categorie sau nu – nu este posibil să trasăm o dreaptă de regresie (precum în regresia liniară).Din păcate, încă o transformare matematică – o transformare logaritmică – este necesară pentru a normaliza distribuirea.

Această transformare logaritmică a valorilor (p) într-o distribuire logaritmică ne permite să creăm o legătură cu ecuaţia de regresie normală.

Distribuţia logaritmică sau transformarea logistică a lui (p) este numită logit (p), care reprezintă raportul probabil că variabila dependentă este 1 şi este definită astfel:

$$logit(p) = log[p/(1-p)] = ln[p/(1-p)] \qquad (1)$$

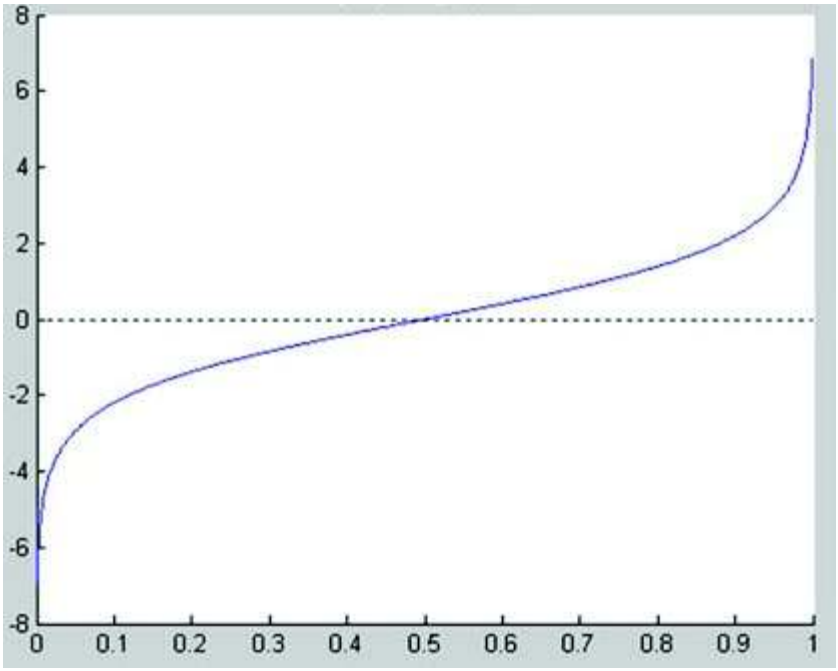


Figura 1. Distribuția logaritmică

Întrucât (p) poate varia numai de la 0 la 1, scala logit (p) variază de la $-\infty$ la ∞ (a se vedea Figura 1.), și este dispusă simetric în jurul funcției logit de 0.5 (care este zero).

Formula (2) prezintă relația dintre ecuația de regresie obișnuită (a+bx +...), care este formula unei linii drepte, și ecuația de regresie logistică.

Formula ecuației de regresie logistică este:

$$\text{logit}[p(x)]=\log\left[\frac{p(x)}{1-p(x)}\right]=a+b_1x_1+b_2x_2+b_3x_3\ldots \qquad (2)$$

Aceasta arată precum o regresie liniară şideşi, regresia logistică găseşte o ecuaţie ce i se potriveşte mai bine, aşa cum face şi regresia liniară, principiile după care se realizează acest lucru sunt diferite.

În loc de folosirea criteriului deviaţiei celor mai mici pătrate pentru cea mai bună potrivire, aceasta utilizează metoda verosimilității maxime (LL), care măreşte la maxim probabilitatea de a primi rezultatele observate, daţi fiind coeficienţii de regresie potriviţi.

O consecinţă a acestui fapt este corectitudinea potrivirii şi statisticile generale de semnificaţie folosite în regresia logistică, fiind diferite de cele folosite în regresia liniară, formula (3), și poate fi privită ca rescrierea formulei (2):

$$p = \frac{exp^{(a+b_1x_1+b_2x_2+b_3x_3+\ldots)}}{1+exp^{(a+b_1x_1+b_2x_2+b_3x_3+\ldots)}} \qquad (3)$$

unde:

p - probabilitatea ca un caz să fie dintr-o anumităcategorie;

exp- baza logaritmilor naturali (e≅ 2.72);

a - constanta din ecuație;

b - coeficientul variabilei predictor.

În concluzie, regresia logistică – include potrivirea unei ecuaţii a formei cu datele:

$$\text{logit (p)} = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots \qquad (4)$$

Interpretarea probabilităţii logaritmice şi a raportului de probabilitate

Coeficienţii logit standardizaţi sunt coeficienţii „b” (valorile pantei) din ecuaţia 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, aşa cum face regresia celor mai mici pătrate (OLS).

Pentru o variabilă dihotomică, şansele de a fi membru al grupului ţintă sunt egale cu probabilitatea de a fi membru în grupul ţintăîmpărţit la probabilitatea de a fi membru în grupul celălalt.

Un concept important îl reprezintă raportul de probabilitate (OR), care estimează schimbările în şansele apartenenţei la grupul ţintă pentru o creştere de o unitate la predictor, fiind calculat prin folosirea unui coeficient de regresie al predicţiei ca fiind exponentul sau exp.

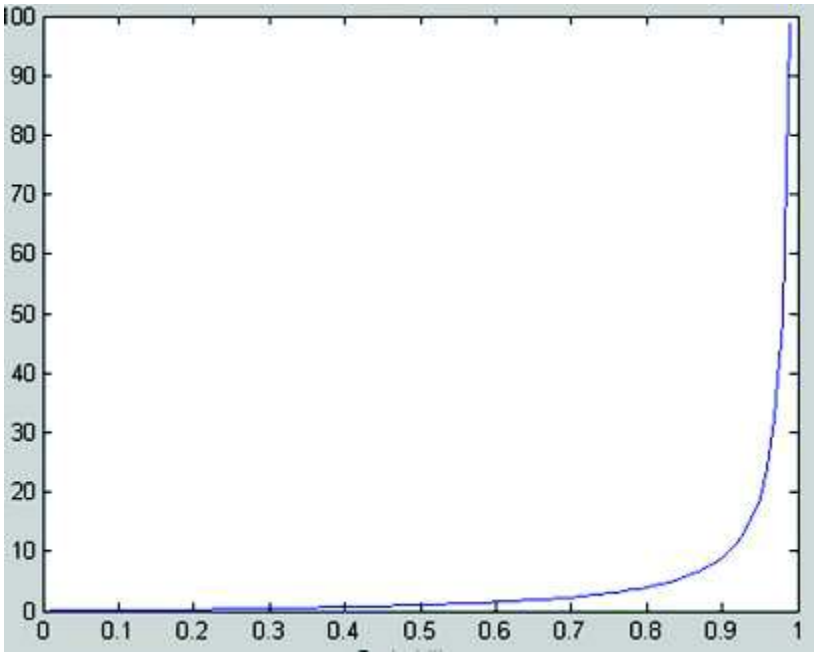


Figura 2.Valoarea raportului de probabilitate

Valoarea raportului de probabilitate (a se vedea Figura 2.) poate varia de la 0 la ∞ şi indică cât de probabil este ca o observaţie să fie mai degrabă parte din grupul ţintă decât membru al celuilalt grup.

Construirea modelului. Verosimilitatea logaritmică

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 şi 1, nu putem folosi metoda celor mai mici pătrate.

Verosimilitatea maximă (ML) este utilizată pentru a afla funcţia care va maximiza abilitatea noastră de a predicta probabilitatea lui Y bazată pe ceea ce ştim despre X. Cu alte cuvinte, (OR) găseşte cele mai bune valori pentru formula (2).

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

În regresia logistică, sunt importante două ipoteze:

- ipoteza nulă, atunci când toţi coeficienţii din ecuaţia de regresie iau valoarea zero;
- ipoteza alternativă care arată că modelul cu predictori ce este luat acum în considerare este precis, şi diferă semnificativ de mulţimea vidă z, de exemplu oferă în mod semnificativ ceva mai mult decât probabilitatea şi nivelul de predicţie 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, şi pentru a fi mai accesibil, este folosit logaritmul natural, producând o verosimilitate logaritmică(LL).

Probabilităţile sunt întotdeauna mai puţin de una, aşa 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 semnificaţieidiferenţei 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.

Semnificaţia la nivelul .05 sau mai scăzut înseamnă că modelul de cercetare cu predictori este diferit în mod semnificativ faţă de cel cu doar o constantă (toţicoeficienţii "b" fiind zero). Aceasta măsoară îmbunătăţirea pe care o fac variabilele explicative în comparaţie cu modelul nul, iar χ^2 este folosit pentru a conferi semnificaţie acestui raport.

Când probabilitatea eşuează să ajungă la nivelul de semnificaţie de 5%, păstrăm ipoteza nulă, ştiind că variabilele independente (predictorul) nu au efecte de creştere (nu fac nicio diferenţă) în a predicta variabila dependentă.

Obținerea parametrilor modelului

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

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

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

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 α și problema noastră va fi, prinurmare, de a obțineparametrulmodelului, β când $y = x \cdot \beta + \varepsilon$.

Daca am făcut n observații, cu y_i răspunsuri și x_i predictori, putem defini:

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

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

Sistemul pe care dorim să-l rezolvăm pentru a găsi parametrul β estescris 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 minimizeazăsumapătratelorrezidurilor.

Regresia logistică multiplă

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:

$$p(\vec{x}) \equiv P(Y = 1 \mid \vec{x}) \tag{5}$$

Deoarece Y este binar, modelarea lui $p(\vec{x})$ este realmente modelarea lui $E(Y \mid \vec{x})$, 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) $\beta_0 + \beta_1 x_1 + ... + \beta_p x_p$, atunci modelul poate duce la probabilități estimate care sunt în afara intervalului [0,1]. Pentru a funcționa mai bine se presupune că:

$$p(\vec{x}) = \frac{exp^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1 + exp^{\beta_0+\beta_1 x_1+...+\beta_p x_p}} \tag{6}$$

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

Este greșit să presupunem că variabilele explicative netransformate sunt tot ceea ce este nevoie vreodată. 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 predictori poate să nu fie adecvat, dar un model de 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:

$$\log\left(\frac{p(\vec{x})}{1 - p(\vec{x})}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{7}$$

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

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

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

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

Bibliografie:

1.Tan, P. Steinbach, M. &Kumar, V., Introductionto Data Mining, PearsonEducation, 2006.

2.Tom M. Mitchell, Generative and discriminative classifiers: Naive Bayesand logistic regression. 2005.

3.P. Komarekand A. Moore, Fast Robust Logistic Regression for Large Sparse Data set swith Binary Outputs, 2003.