A logisztikus regresszió és az általánosított lineáris modell

Ferenci Tamás
tamas.ferenci@medstat.hu

2018. február 9.

Tartalom

Kvalitatív változó eredményváltozó pozíciójában

- Például a feladat egy „csődbe megy-e vagy sem” jellegű változó modellezése
- Ez bináris változó $\rightarrow$ mint az eddig tárgyalt dummy változók, csak ezúttal eredményváltozóként
- Jelent ez módosulást? (Hiszen például magyarázó változóként mindegy volt, hogy egy változó bináris, az OLS-t nem zavarta, hogy történetesen csak 0 és 1 értékeket vesz csak fel)
- Most drasztikusan más a helyzet: *Y nem modellezhető* OLS-sel

OLS és a bináris eredményváltozó

- Matematikai részletekbe nem megyünk bele
- Intuitíve: gondoljunk arra, hogy az OLS – elvileg – bármilyen értéket becsülhet $-\infty$ és ∞ között $\rightarrow$ egy ilyen hogyan lenne értelmezhető egy „csődbe megy-e vagy sem" kérdés válaszaként?!
- De: mégis lineáris struktúrában fogjuk megoldani a problémát... csak trükkösebben alkalmazzuk: bináris Y helyett egy transzformált változóra
- Avagy – fordítva megfogalmazva – megtartjuk a lineáris kombinációt, de annak az eredményét áteresztjük egy olyan függvényen, ami a $(-\infty, \infty)$-t a $[0, 1]$-re képezi le

A mostani feladat általánosabban

- Tegyük fel, hogy elkészült a bináris Y-ra adott modellünk, és azt előrejelzésre használjuk
- Vegyük észre, hogy az Y szerinti érték egyfajta *csoporttagságot* jelent: becsődölő, működő
- Az előrejelzés ebben a kontextusban lényegében besorolás egy csoportba!
- Tehát mégegyszer: a megfigyelési egység két csoport valamelyikébe tartozik, mi a csoporttagságával összefüggő adatok alapján tippeljük meg a csoporttagságot
- Ezt a feladatot általában *osztályozásnak* (klasszifikáció) nevezik
- A klasszifikáció hatalmas gyakorlati jelentőségű feladat: melyik cég megy csődbe (a mérlegadatai alapján), melyik beteg fog meghalni (a laboreredmények alapján), kit vesznek fel adott munkahelyre (egyéni jellemzők alapján) stb. stb.

A feladat átalakítása

- Hogy a kérdést a magyarázó változók lineáris kombinációjával tudjuk kezelni, áttérünk más változóra
- Először is: nem az 1-es csoportba tartozás tényét, hanem annak $\mathbb{P}_{\underline{X}}$ feltételes valószínűségét fogjuk modellezni
- Az alsó index értelme: az 1-es csoportba tartozás valószínűsége, *feltéve*, hogy a magyarázó változók $\underline{X}$ értékűek, azaz precízen: $\mathbb{P}_{\underline{X}} = \mathbb{P}\left(Y = 1 | \underline{X}\right)$
- Ezzel a $\{0, 1\}$ változó helyett egy $[0, 1]$-on lévőt kell modellezni
- Vegyük észre, hogy ezzel még nem léptünk ki az eddigi regressziós keretből, sőt, teljesen megfelelünk neki, hiszen egy bináris (0-1) változóra ez a feltételes valószínűség épp a feltételes várható érték!
- Azt fogjuk mondani, ez a későbbiek szempontjából lesz fontos, hogy az eredményváltozó eloszlása Bernoulli (p valószínűséggel vesz fel 1-et, $1 - p$ valószínűséggel 0-t), és ennek a feltételes várható értékét modellezzük

A feladat további átalakítása

- Ez persze még mindig kevés, ezért újabb transzformációt alkalmazunk
- Odds (esély) fogalma: az 1-es csoportba tartozás valószínűsége a 0-s csoportba tartozás valószínűségéhez viszonyítva, jelen esetben valószínűség osztva 1-valószínűséggel
- Azaz

$$\text{odds}_{\underline{X}} = \frac{\mathbb{P}_{\underline{X}}}{1 - \mathbb{P}_{\underline{X}}}$$

- Könnyen megoldható $\mathbb{P}_{\underline{X}}$-re:

$$\mathbb{P}_{\underline{X}} = \frac{\text{odds}_{\underline{X}}}{1 + \text{odds}_{\underline{X}}}$$

És még egy átalakítás

- Az odds már a $[0, \infty)$ intervallumon van
- Majdnem jó, egy utolsó trükk: bevezetjük a *logit* fogalmát, mint log-odds:

$$\text{logit}_{\underline{X}} = \ln \text{odds}_{\underline{X}}$$

- És ez már a $(-\infty, \infty)$-n van (és szimmetrikussá is tettük a siker és kudarc eloszlását rajta)!
- Na, ezt fogjuk lineáris struktúrával modellezni!

$$\text{logit}_{\underline{X}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k = \underline{X}^T \boldsymbol{\beta}$$

- A módszer neve: logit regresszió, vagy logisztikus regresszió

A logisztikus regresszió visszafejtése

- Játszuk el mindezt visszafelé, feltéve, hogy β-k már ismert:

$$\text{logit}_{\underline{X}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

$$\text{odds}_{\underline{X}} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k}$$

$$\mathbb{P}_{\underline{X}} = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k}} = \frac{e^{\underline{X}^T\beta}}{1+e^{\underline{X}^T\beta}}$$

- Az utolsó lépésben kapott $f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$ épp a korábban emlegetett, $(-\infty, \infty)$-t a $[0, 1]$-be képező függvény!
- β ismeretében egyszerű algebrai műveletekkel kapjuk a siker valószínűségeit
- És az utolsó lépés: hogy becsüljük meg β-t?
- Sajnos az OLS – ahogy már mondtuk – nem jó, új módszer kell: maximum likelihood (ML) becslés

A logisztikus regressziós modell becslése

- Minden **b** választáshoz meghatározható a minta (itt: adatbázisunk) likelihood-ja (precízen: adott **b** mellett mekkora likelihood-dal jött volna ki a mintánk)
- Ezt fogjuk a **b**-ban maximalizálni, és így kapjuk $\widehat{\beta}_{\mathrm{ML}}$-t
- Kérdés: hogyan kapjuk a minta likelihood-ját?
- Annyira nem nehéz, hiszen *egy* mintaelemre a kijövetelének valószínűsége $\mathbb{P}_{\underline{X}}$ (ha az eredményváltozója 1) illetve $1 - \mathbb{P}_{\underline{X}}$ (ha eredményváltozója 0), mely értékek kiszámíthatóak adott **b** mellett (már láttuk is)
- Már csak az egész mintára (nem egyes mintaelemekre) kell kiszámítani, itt függetlenség feltételezésével élünk

A logisztikus regressziós modell becslése

- Az egész minta likelihood-ja így:

$$L(b_0, b_1, \ldots, b_k) = \prod_{Y_i=1} \mathbb{P}_{\underline{X}_i} \prod_{Y_i=0} \left(1 - \mathbb{P}_{\underline{X}_i}\right) = \prod_{i=1}^{n} \mathbb{P}_{\underline{X}_i}^{Y_i} \left(1 - \mathbb{P}_{\underline{X}_i}\right)^{1-Y_i} =$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\underline{X}_i^T \mathbf{b}}}{1 + e^{\underline{X}_i^T \mathbf{b}}}\right)^{Y_i} \left[1 - \left(\frac{e^{\underline{X}_i^T \mathbf{b}}}{1 + e^{\underline{X}_i^T \mathbf{b}}}\right)\right]^{1-Y_i}$$

- Ezzel a megoldandó feladat:

$$\max_{b_0, b_1, \ldots, b_k} L(b_0, b_1, \ldots, b_k)$$

- E helyett a gyakorlatban inkább a vele ekvivalens

$$\min_{b_0, b_1, \ldots, b_k} -2 \ln L(b_0, b_1, \ldots, b_k)$$

  feladatot oldjuk meg (nem csak numerikus okokból)
- Ennek megoldása szolgáltatja a paraméterek becslését

Elemzés

- Értelmezzük az együtthatókat:

$$\frac{\text{odds}_{x_1,\ldots,x_{l-1},x_l+1,x_{l+1},\ldots,x_k}}{\text{odds}_{x_1,\ldots,x_{l-1},x_l,x_{l+1},\ldots,x_k}} = \frac{e^{x_1,\ldots,x_{l-1},x_l+1,x_{l+1},\ldots,x_k}}{e^{x_1,\ldots,x_{l-1},x_l,x_{l+1},\ldots,x_k}} = e^{\beta_l}$$

- Ezért az e^{β_l}-kat is meg szokták adni a programok, a nevük esélyhányados (odds ratio, OR)

Előrejelzés

- Még egy megfontolást kell tenni: csak csődvalószínűséget kaptunk... de az előrejelzésben konkrét kimenet kell! Mikor soroljuk becsődölőbe? Ha ez a valószínűség 0,5-nél nagyobb? 0,1-nél? 0,99-nél...?
- Jelölje ezt a határt C (cut-off point, cut value):

$$\widehat{Y} = 1 \Leftrightarrow \mathbb{P}_{\underline{X}} > C$$

- Ekkor különböző C-khez különböző konkrét klasszifikációk tartoznak

A klasszifikáció jóságának mérése

- Legalapvetőbb eszköz a klasszifikációs mátrix:

	$\widehat{y}=1$	$\widehat{y}=0$
$y=1$	6	1
$y=0$	5	38

- Főátlóban a helyes osztályozások, ezek aránya a helyes osztályozási ráta (itt $\frac{6+38}{6+1+5+38}=0{,}88$)
- Mellékátlóban: első- és másodfajú hibák (specificitás, szenzitivitás)
- Gondoljuk végig, hogyan változik ezek aránya C növelésére, ill. csökkentésére
- Szenzitivitás az (1-specificitás) függvényeben különböző C-kre: ROC-görbe (terület alatta: AUC)

C megválasztása veszteség-függvény alapján

- Ha tudjuk, hogy az egyes hibák milyen „költséget” jelentenek, akkor analitikusan választhatunk optimális C-t
- Veszteség-mátrix:

	$\widehat{y} = 1$	$\widehat{y} = 0$
$y = 1$	0	1
$y = 0$	0,2	-0,2

- Ezzel az előző klasszifikációs mátrix költsége:

$$6 \cdot 0 + 1 \cdot 1 + 5 \cdot 0{,}2 + 38 \cdot (-0{,}2) = -5{,}6$$

- Azt a C-t választjuk, aminél ez minimális!

C megválasztása veszteség-függvény nélkül

- C korrekt megválasztása *csak* veszteség-függvény ismeretében lehetséges: ha nem tudjuk, hogy milyen súlyú a kétféle hibázás, akkor honnan tudhatnánk egyáltalán megmondani, hogy mi az, hogy „jó” választás?
- Néha azonban mégis rákényszerülünk a veszteségek ismerete nélküli döntésre
- Klasszikus (nem ROC-görbére támaszkodó) heurisztikák:
  - Fix 0,5-ös cutoff
  - A cutoff legyen az 1-esek mintabeli aránya
  - A cutoff legyen olyan, hogy azzal a predikált 1-esek aránya megegyezzen az 1-esek mintabeli arányával
- Optimalizálás a ROC-görbe alapján:
  - A specificitás és a szenzitivitás összege legyen maximális (Youden-szabály)
  - A bal felső – optimális – ponthoz legközelebbi pont választása (azaz $(1 - Se)^2 + (1 - Sp)^2$ legyen minimális)

Modelljellemzés pszeudo-R^2 mutatóval

- Az OLS-nél látott R^2-hez hasonló elvű („hol járunk az úton?”) mutató szeretnénk LR-re is
- Az *ESS* helyett itt a $-2 \ln L$ jellemzi a modellt
- Mi a tökéletes modell? $\rightarrow \mathbb{P}_{\underline{X}} = 1$ ha $Y = 1$ és $\mathbb{P}_{\underline{X}} = 0$ ha $Y = 0 \rightarrow$ mennyi ennek a likelihoodja?
- Épp 1, $-2 \ln L = 0$
- Az üres – semmilyen magyarázó változót nem tartalmazó modell – $-2 \ln L$-je analitikusan meghatározható (analóg a helyzet az OLS-sel)
- Az alapján a McFadden-féle pszeudo-R^2:

$$R^2 = \frac{\left(-2 \ln L_{\text{null}}\right) - \left(-2 \ln L_{\text{targy}}\right)}{-2 \ln L_{\text{null}}}$$

- Sok fenntartás van az ilyen mutatókkal kapcsolatban!

Modellszelekció

- Nested modellszelekció,

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_{q+m} = 0$$

- Ha nagy mintánk van, akkor rendkívül kényelmesen vizsgálható egy új próbakészítési elvvel, az ún. likelihood-hányados (LR) elven konstruált teszttel:

$$\left(-2 \ln \widehat{L}_{H_0}\right) - \left(-2 \ln \widehat{L}_{H_1}\right) \sim \chi^2_m$$

- Üres modelltől való szignifikáns különbözés tesztelése: függetlenségvizsgálat
- Szaturált modelltől van szignifikáns különbözés tesztelése: illeszkedésvizsgálat

A lineáris és a logisztikus regresszió közös keretben

- Vegyük észre a hasonlóságokat!
  1. Van valamilyen eredményváltozó-eloszlás
     - Lineárisnál normális, logisztikusnál Bernoulli
  2. A feltételes várhatóérték valamilyen transzformáltját modellezzük:
     $g\left[\mathbb{E}\left(Y|\underline{X}\right)\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$
     - Lineárisnál az identitás, logisztikusnál a korábban látott f (pontosabban szólva annak az inverze)
  3. Elvileg valamit mondani kellhet a varianciáról is
     - Lineárisnál azt, hogy $\sigma_i^2 = \sigma_0^2$, logisztikusnál megspóroltuk ezt, mert a várható értéke meghatározta a szórást is (egy paramétere volt az eloszlásnak)

Az általánosított lineáris modell (GLM)

- A fenti komponensek határozzák meg az ún. általánosított lineáris modellt (generalized linear modell, GLM)
- Az eredményváltozó eloszlása legyen exponenciális eloszláscsaládból származó
- A g függvény neve: link függvény
- Becslés maximum likelihood-dal
- A lineáris és logisztikus regresszió mind speciális esete ennek (alkalmasan választott eredményváltozó eloszlással, link függvénnyel és szórás-függvénnyel)
- Sok minden más is ide tartozik, lássunk még egy példát

Poisson regresszió

- Mi van, ha az eredményváltozó valamilyen darabszám, események száma jellegű változó (count data)?
- Ilyenekre tipikusan feltételezett eloszlás első közelítésben: Poisson-eloszlás
- Ez exponenciális családbeli
- Várható értéke itt is épp a paramétere
- Tipikus link függvény választás: a log
- Összerakva mindezeket a modellünk:

$$Y \sim \mathrm{Poi}(\lambda)$$

$$\log\left[\mathbb{E}\left(Y|\underline{X}\right)\right] = \log\lambda = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$