

PSY252

Statistická analýza dat v psychologii II

Seminář 5-6

Logistická regrese

Logistic regression

Předpovídáme pohlaví pachatele

Víme, že pachatel nosí náušnici/e a napsal dopis se skórem emočních adjektiv 8.

Víme, že...

- náušnice nosí 24% mužů a 86% žen
- na škále přítomnosti emočních adjektiv od 1 do 13 mají ženy průměr 9,1 a muži pouze 4,5.

Jaká je pravděpodobnost, že pachatel je žena?

Logistická regrese

- Rozšíření lineární regrese na dichotomické závislé
 - není to lineární regrese, protože nejde o lineární vztah
 - Závislou kódujeme 1 (jev nastal) a 0 (jev nenastal)
 - Ideově je závislou proměnnou pravděpodobnost toho, že jev nastal(nastane)
 - Technicky je závislou proměnnou **šance**
 - Pomocí prediktorů predikujeme, jaká je šance, že jev nastane.
-

Technický základ logistické regrese 1

- šance $O_{Y=1} = P_{Y=1}/P_{Y \neq 1} = P_{Y=1}/(1-P_{Y=1})$
- $\ln O_{Y=1}$ se jmenuje **logit** ($P_{Y=1}$)
- 2 ekvivalentní rovnice logistické regrese

$$\ln O_{Y=1} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_m X_m$$

$$P_{Y=1} = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + \dots + b_m X_m)}}$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	nausnice	2,147	1,144	3,523	1	,061	8,556	,909	80,501
	emoce	,387	,199	3,791	1	,052	1,472	,997	2,173
	Constant	-3,797	1,545	6,041	1	,014	,022		

a. Variable(s) entered on step 1: emoce.

$$P_{Y=\text{žena}} = \frac{1}{1 + e^{-(-3,8 + 0,4EM + 2,1NA)}}$$

$$\ln O_{Y=\text{žena}} = -3,80 + 0,39\text{emoce} + 2,15\text{náušnice}$$

- Pro náušnice=1 a emoce=8 ... $P=0,81$ $O=4,2$
- Kdyby neměl náušnici ... $P=0,33$ $O=0,50$
- Změna náušnice z 1 na 0 způsobila
8,5násobný pokles šancí ... e^b

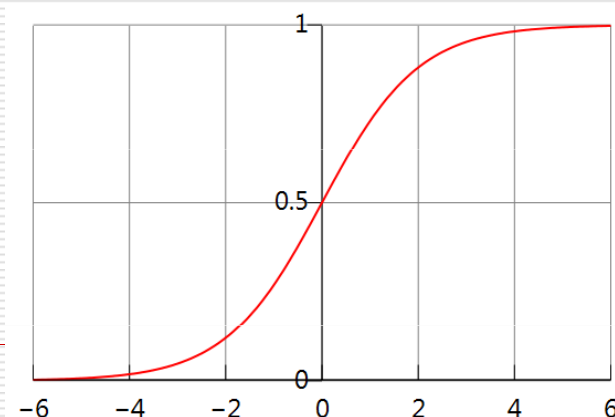
Proč tak složitě?

Závislá jako pravděpodobnost má měřítko v rozsahu $\langle 0;1 \rangle$. Kombinace prediktorů má ale rozsah $(-\infty; \infty)$.

Proto změníme měřítko závislé

1. Místo P použijeme O s měřítkem $\langle 0; \infty \rangle$
2. Pomocí logaritmu změníme měřítko na $(-\infty; \infty)$.

Také lze říci, že jde o linearizaci vztahu.



Technický základ logistické regrese 2

Jak spočítáme regresní váhy, které vyústí v nejlepší predikci pravděpodobnosti $Y=1$?

- nespočítáme, odhadneme (zapomeňme na nejmenší čtverce)
 - odhad metodou **maximální věrohodnosti** (maximum-likelihood estimation)
 - Výpočetně složitý algoritmus
 - Dochází k takovým váhám, s nimiž je podmíněná pravděpodobnost získání dat, která jsme získali, nejvyšší možná : $P(\text{data} | b_0, b_1, \dots, b_m) = \max$
 - likelihood = jiné slovo pro podmíněnou p-nost
-

Jak dobře regrese predikuje?

- Likelihood je měřítkem zdařilosti regrese v logaritmované podobě: **log-likelihood**

$$LL = \sum_{i=1}^N [Y_i \ln P_{Y=1} + (1 - Y_i) \ln(1 - P_{Y=1})]$$

- **LL** sumíruje shodu mezi odhadem a daty
 - maximem je 0, minimem je $-\infty$
 - častěji se udává jako **-2LL**, tj. vynásobený -2
-

Predikuje regrese lépe než *nic*?

- nic = základní model (baseline model) = predikujeme všem 0 nebo 1, podle toho, co z toho se vyskytuje častěji = $P_{Y=1}$ je pro všechny lidi stejná
 - Potom můžeme srovnat model s prediktory s tímto základním modelem.
 - rozdíl $-2LL$ obou modelů má χ^2 rozložení s df =počet prediktorů
$$\chi^2 = -2LL_{\text{náš model}} - -2LL_{\text{základní model}}$$
$$df = m_{\text{náš model}} - m_{\text{základní model}}$$
 - Podobně můžeme srovnávat i modely s různým počtem prediktorů mezi sebou
-

Nedalo by se to trochu zjednoduřit?

-2LL lze převést na ukazatele podobné R^2

R_L^2 Hosmera a Lemeshowa

R_{CS}^2 Coxe a Snella

R_N^2 Nagelkerkeho

Nabývají hodnot od 0 do 1.

Udávají jak moc díky prediktorům klesl -2LL

Praktické problémy

- Regresní koeficienty se nevypočítávají, ale iteračně odhadují.
 - Iterace nemusí vždy proběhnout úspěšně
 - nemusí konvergovat
 - mohou se vyskytnout bláznivé hodnoty
 - Problematické výsledky naznačují nedostatky v datech
 - při absenci některé z kombinace hodnot prediktorů a závislé
 - při dokonalé predikci
 - LR je náročná na velikost vzorku
-