Logisztikus regresszió

Bekövetkezés esélye

- Valószínűség (P): 0 és 1 közötti valós szám, az esemény bekövetkezésének esélyét fejezi ki. Fej dobásának esélye: 1:2 $= \frac{1}{2} = 0{,}5$.
- Odds/esélyérték (O): a tét hányszorosa lesz a nyeremény, vagyis a nyerés esélye. Azaz: hányszor akkora a valószínűsége annak, hogy valami bekövetkezik, mint az, hogy nem. Ha fejre fogadok, a nyerés esélye: 1:1 $= \frac{1}{1} = 1$. Ha vesztek, 1-et vesztek, ha nyerek, 1-et nyerek. Értékek: 0 és ∞ között. Előnye: tartalmazza a megfigyelések számát.
- Logit (L): az odds értékének *e*-alapú logaritmusa. Értéke 1-es odds-ra 0, 3-asra 1,99, 0,33-ra $-1{,}99$. A szélső értékei $-\infty$ és $-\infty$. Előnye: nagyobb számértéket kisebb számmal lehet kifejezni.

Matematikai összefüggések

P	0	0,01	0,1	0,5	0,9	0,99	1
O	0	0,0101	0,111	1	9	99	∞
L	$-\infty$	$-4,60$	$-2,20$	0	2,20	4,60	∞

$O = \frac{P}{1-P}$, $P = \frac{O}{1+O}$

$L = ln(O) = ln(\frac{P}{1-P})$, ami az Euler-féle számot veszi bázisul.

R-ben 2-es logaritmus: `log2()`, 10-es: `log10()`, e-alapú `log()`.

Logit vagy log odds: a valószínűségi érték transzformálása úgy, hogy bármilyen értéket felvehessen, ne csak 0 és 1 közöttit.

A logisztikus modellben nincs hibaterminus és variancia.

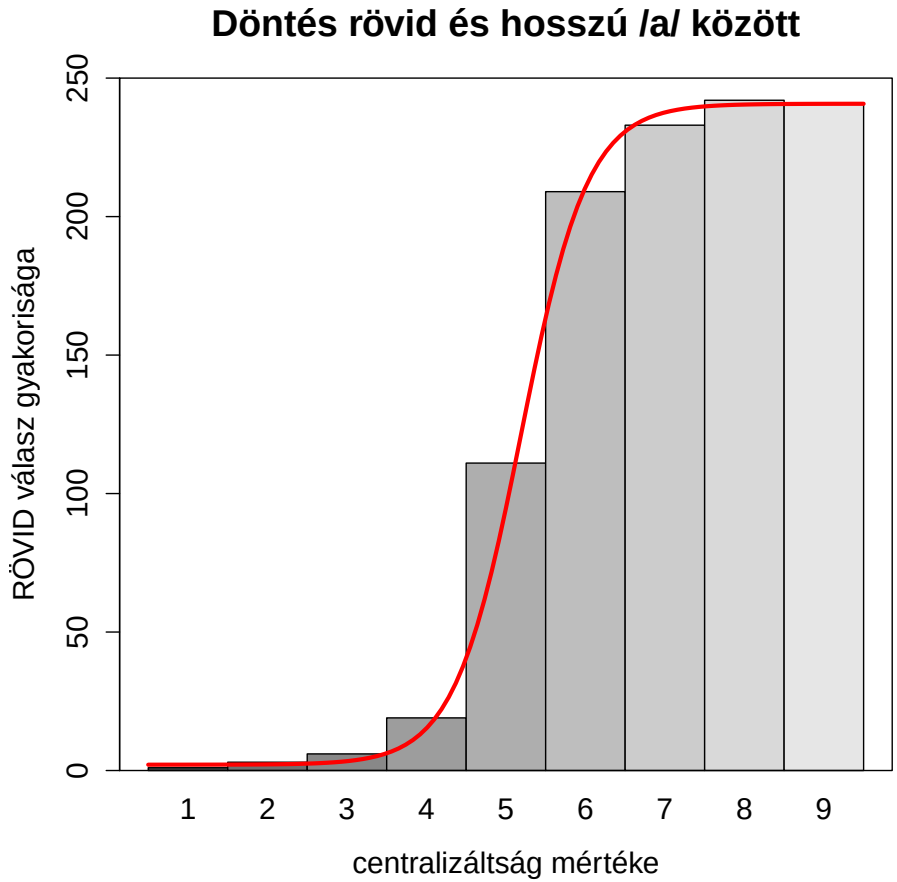
Logisztikus regressziós modell paremétereinek becslése: legnagyobb valószínűség (*maximum likelihood*), hasonlóan a lineáris regresszió legkisebb négyzetek által meghatározott regressziós egyenes becsléséhez. Itt: azon paraméterek megtalálása, amelyek mellett legvalószínűbb, hogy éppen a megfigyelt értékeket kapjuk.

Illeszkedés jóságát adja meg, azaz az egyes adatpontok összes valószínűsége.

Generalised linear models: a modell értékeinek visszavezetése lineáris értékekre egy összekötő függvényen keresztül. Itt a függvény $L = ln(\frac{P}{1-P})$.

Logisztikus függvény

Példa: magyar *á* és *a* magánhangzó közötti átmenet az 1. és 2. formáns távolságának lépésenkénti változtatásával. Kérdés: 1. hol van a kategóriahatár a két hang között, 2. milyen éles a kategóriahatár?



Lehetséges alkalmazások:

- Táblázatba rendezett adatok gyakoriságokkal és binomiális adatokkal, pl. horkoló, dohányzó és túlsúlyos személyek között mekkora arányban fordul elő magas vérnyomás, szemben a nem horkoló, de dohányzó és túlsúlyos személyekkel stb. `glm(...,family="binomial")` táblázatban összefoglalt adatokra.
- Bináris döntések, pl. szómemorizálási feladat szófajok szerint: előfordult-e egy adott szó egy adott szövegben. `lrm()` az `rms` csomagban, ha soronként egy megfigyelésünk van.
- Kevert modellek alkalmazása manipulált körülmények között, például *mész* – *méz* döntés, ha a frikatíva zöngésségét 0 és 100% között manipuláljuk 11 lépésben. `glmer(..., family="binomial")`, `lme4` csomag.

Példák: logreg.txt

summary(glm())

Deviance residuals: elvárt megfigyelésektől való eltérés pozitív és negatív irányba, hasonlóan a reziduálisokhoz a lineáris modelleknél. Minél nagyobb az eltérés, annál gyengébb a modell illeszkedése.

`Dispersion parameter for binomial family taken to be 1`: a logisztikus regressziós modell nem tartalmazza a varianciát, hiszen cellánként egy értékünk van.

Residual deviance: egy χ^2 eloszlásra illesztett érték, 4-es szabadsági fokra 9,49-es határértékkel 5%-os konfidenciahatár esetén, tehát a modell jósága bőven megfelelő.

`Number of Fisher Scoring iterations: 4`: modellillesztések száma, amik után a jelenlegi output létrejött. Default maximum: 25.

Faktorhatások értelmezése

`summary(h,corr=T)`: ha az egyes faktorok közötti korreláció alacsony, a nélkülük számolt modell nem térne el szignifikánsan a jelenlegitől.

Mivel a z-érték alapján a dohányzás hatása nem szignifikáns, lehet vele egyszerűsíteni a modellt.

Példa

Baayen 2008, *Logistic regression* c. fejezet, `languageR` csomag, `english` adatmátrix.

Lexikális döntés: a képen látható alak létező szó-e?
`english$CorrectLexdex`: 30-ból hány ember azonosította a szót létező szóként.

Milyen nyelvi kategóriák befolyásolják a szófelismerést? Hogyan függenek össze a felismerési adatok a `RTlexdec` változóban tárolt reakcióidőkkel?

GLM soronkénti adatokra

Ha nem gyakorisági táblázatokkal dolgozunk, hanem egy adat = egy sor: `lrm()` függvény a `rms` csomagból.

Baayen példája: `regularity` adatmátrix a `languageR` csomagból. Holland szavak szabályos és szabálytalan ragozása és az ezt befolyásoló potenciális faktorok (gyakoriság, valencia stb.).

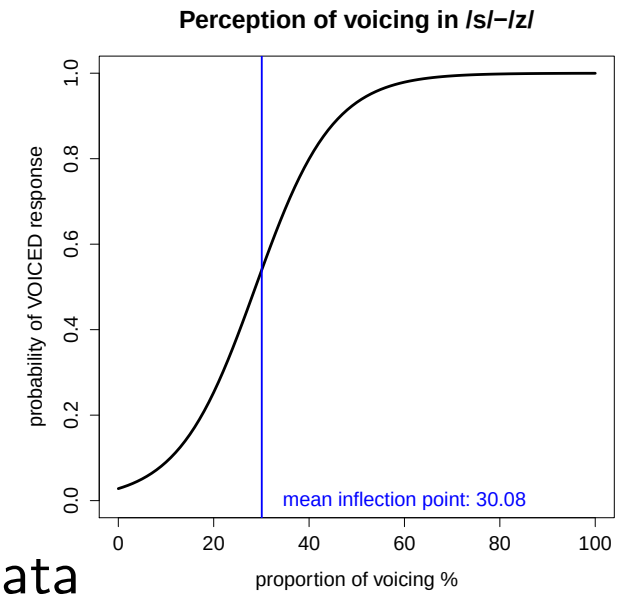
```
h =
lrm(Regularity~InflectionalEntropy+Valency,regularity)
```

Eredmények megtekintése NEM `summary()` függvénnyel, hanem `anova()`-val vagy modellobjektum nevének beírásával, itt `h`.

Generalised linear mixed models

Logisztikus regresszió számítása bináris vagy kategoriális ($k = 2$) adatokra random hatással.

Összehasonlítás alapja a lineáris kevert modellekhez hasonlóan: intercept (k) és meredekség (m) és ennek alapján inflexiós pont ($-k/m$) minden egyes random hatásként definiált egységre (beszélő, item stb.).



Adatok: devoice.RData

```
h = glmer(response∼prop.voice+(1+prop.voice|subj),
family="binomial",data=devoice)
```

Görbe ábrázolása a `coef(h)` függvényből kinyert k és m együtthatók alapján. Összes eredmény ábrázolása átlagolással.

```
curve(exp(mean(d.coef$m)*x+mean(d.coef$k))
+(1+exp(mean(d.coef$m)*x+mean(d.coef$k)))
+xlim=c(0,100),ylim=c(0,1))
```