Logisztikus regresszió

Függő változó típusai egy vagy több független változó esetén

- (Para)metrikus: intervallum- vagy arányskála. Modellezés leggyakoribb módja: lineáris regresszió.
- Ordinális: oridnális skála. Ha nincs okunk metrikus skálának tekinteni, Cumulative Link Models (CLM).
- Nominális: eddig csak a khí-négyzet próbával találkoztunk. Hátrány: legfeljebb két csoportot tudunk összehasonlítani, és ismételt mérésekre sem alkalmas.

Megoldás: általánosított lineáris modellek

Lényegük: az értékeket egy függvényhez rendelik, amin keresztül visszavezethetők egy lináris függvényre.

Legismertebb eloszlások: binomiális eloszlás, Poisson-eloszlás.

A bináris kategoriális változók modellezésének legelterjedtebb módja a binomiális vagy szigmoid függvény.

Alkalmazás: két opciós feleletválasztós tesztek (jó/rossz, igen/nem). Általánosabban: egy esemény bekövetkezik, vagy sem. Pl. három korosztályban a dohányzás magas vérnyomáshoz vezet, vagy sem.

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.

Esélyhányados (odds ratio)

A soproniak és a győriek italpreferenciáját mérik fel. Néhány órán keresztül megkérdeznek minden szembejövőt, hogy mit iszik gyakrabban: bort vagy sört? Az *egyiket sem* válaszokkal nem foglalkoznak.

	bor	sör
Sopron	184	15
Győr	91	113

Esélyhányados (*odds ratio*):

A megkérdezettek 2. csoportja (Győr) esetében a válaszok 2. faktorszintje / 1. faktorszint, itt: 1,241.

A megkérdezettek 1. csoportja (Sopron) esetében a válaszok 2. faktorszintje / 1. faktorszint, itt: 0,081.

A 2. sor aránya / első sor aránya: 15,32.

Esélyhányados értelmezése

- Ha a hányados nagyobb 1-nél, a 2. faktorszint (itt: sör) jellemzőbb (= nagyobb hatással van) a 2. csoportra (itt: Győr).
- Ha a hányados kisebb 1-nél, a 2. faktorszint kevésbé jellemzőbb (= kisebb hatással van) a 2. csoportra.
- Ha a hányados 1, a vizsgált változónak (italpreferencia) nincs hatása.

Esélyhányados hátránya: 1 fölött bármilyen értéket felvehet, de 1 alatt 0 a határa.

Kiút: az esélyhányados logaritmusával számolunk. 0: a független változónak nincs hatása, pozitív szám: a 2. faktorszint nagyobb hatással van a 2. csoportra, negatív szám: a 2. faktorszint kisebb hatással van a 2. csoportra.

Összefüggés a valószínűséggel

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 e 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.

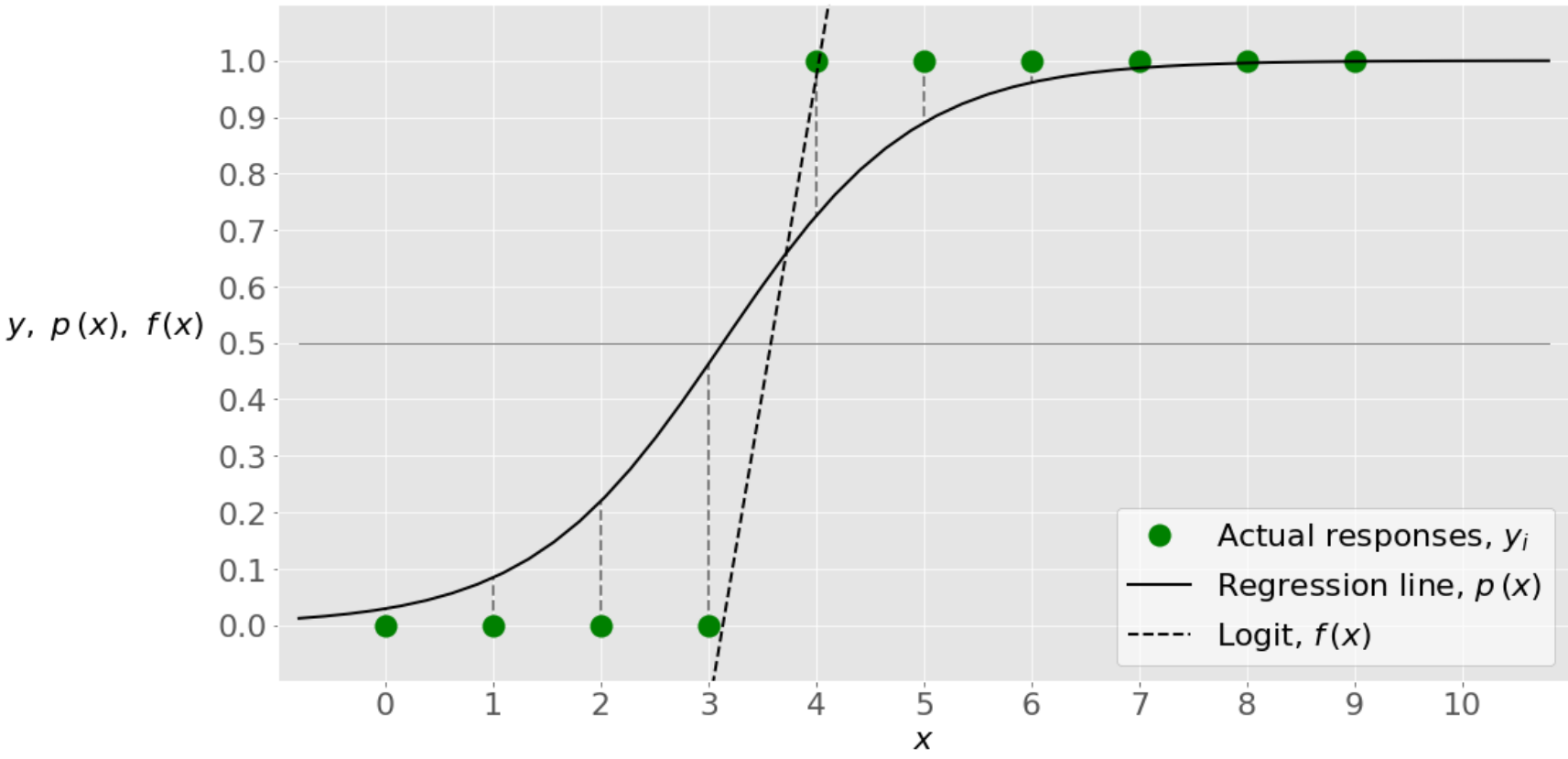
A logisztikus modellben nincs hibaterminus és variancia.

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.

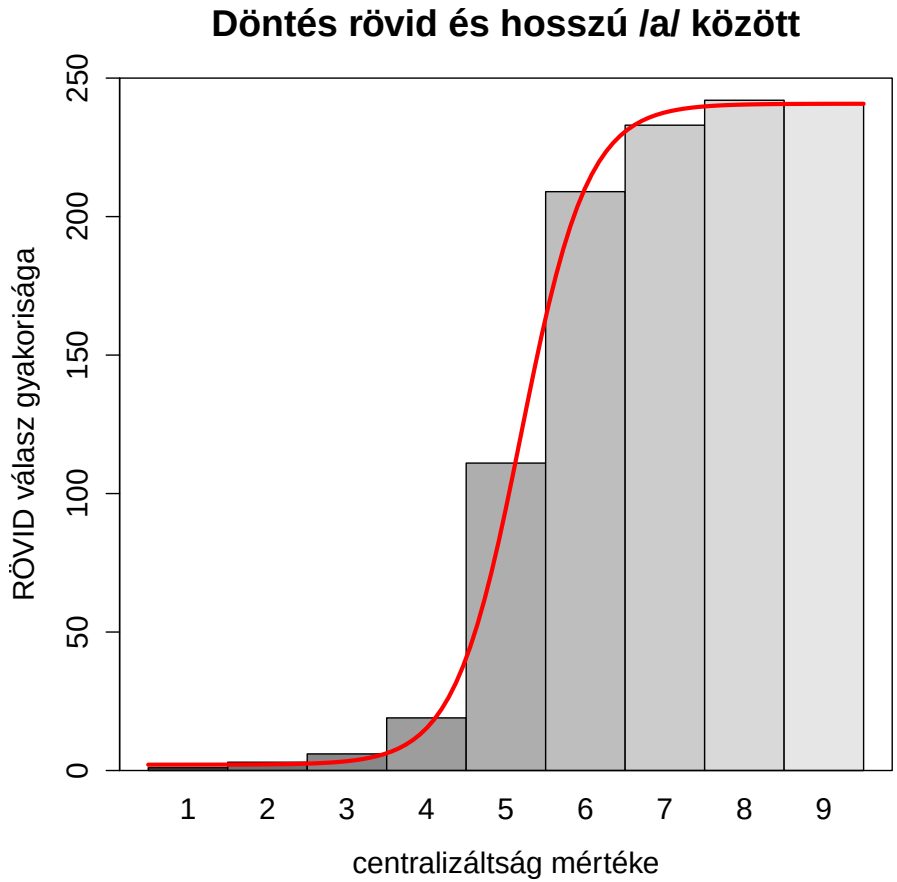
Generalised linear models: a modell értékeinek visszavezetése lineáris értékekre egy összekötő függvényen keresztül. Itt a függvény $L = ln(\frac{P}{1-P})$.

Logisztikus függvény



Logisztikus függvény

Példa: magyar *á* és *a* magánhangzó közötti átmenet az 1. és 2. formáns távolságának lépésenkénti csökkentésével. 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, például *mész* – *méz* döntés, ha a frikatíva zöngésségét 0 és 100% között manipuláljuk 11 lépésben, vagy bináris feleletválasztós teszt. `glmer(..., family="binomial")`, `lme4` csomag.

summary(glm())

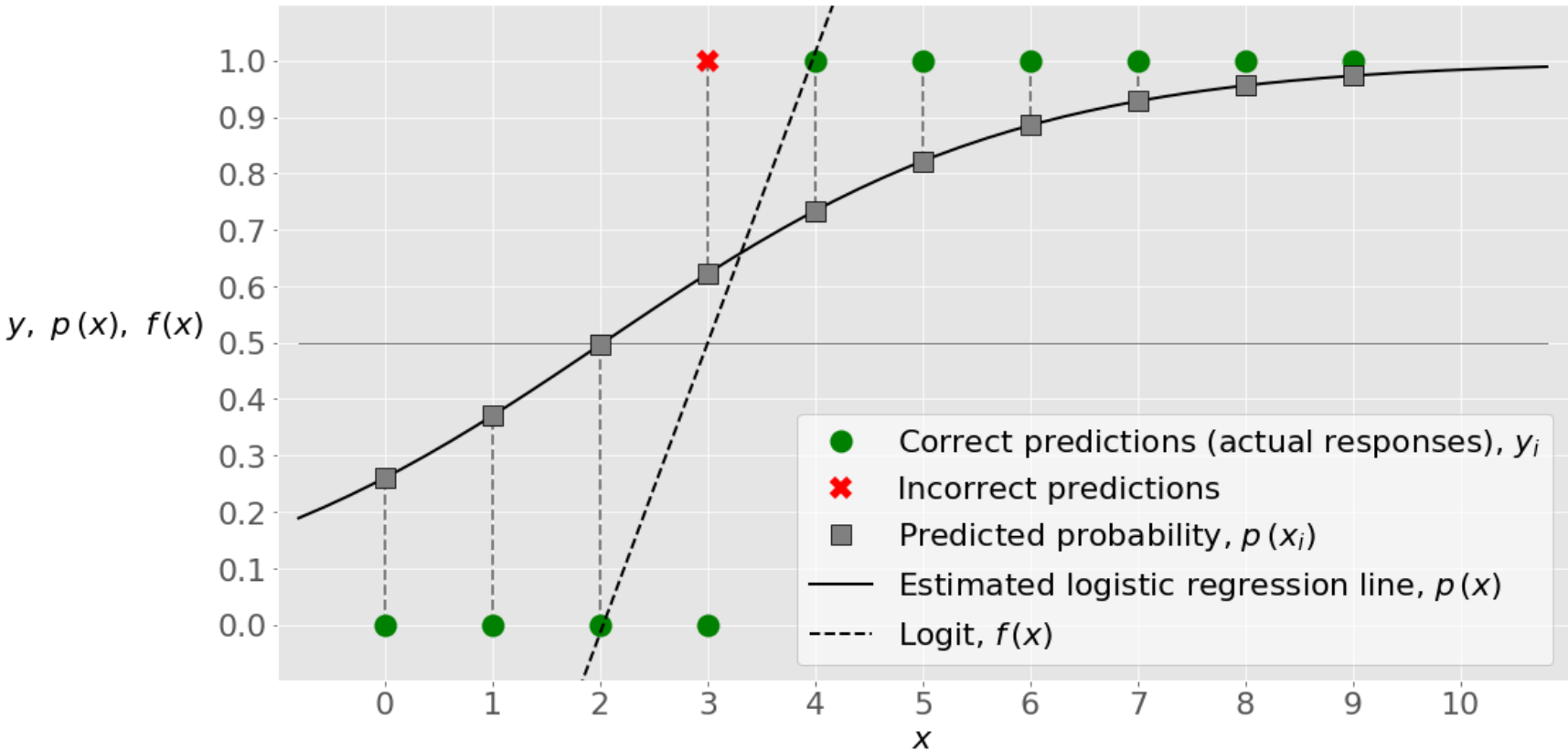
Példák: logreg.txt
Intercept: a binomiális görbe metszéspontja.
Estimate: ha pozitív, a második faktorszint magasabb gyakoriságokhoz vezet, vagyis a szigmoidgörbe pozitív irányba nő.
`summary(h,corr=T)`: ha az egyes faktorok közötti korreláció alacsony, a nélkülük számolt modell nem térne el szignifikánsan a jelenlegitől.

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

Reziduális eltérés:



Példa

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

Lexikális döntés: a képen látható alak létező szó-e?
`english$CorrectLexdec`: 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`.