TÖBBVÁLTOZÓS ADATELEMZÉS

Logit modell

2020.10.05.

© Dr. Vékás Péter, e-mail: peter.vekas@uni-corvinus.hu
BCE Matematikai és Statisztikai Modellezés Intézet

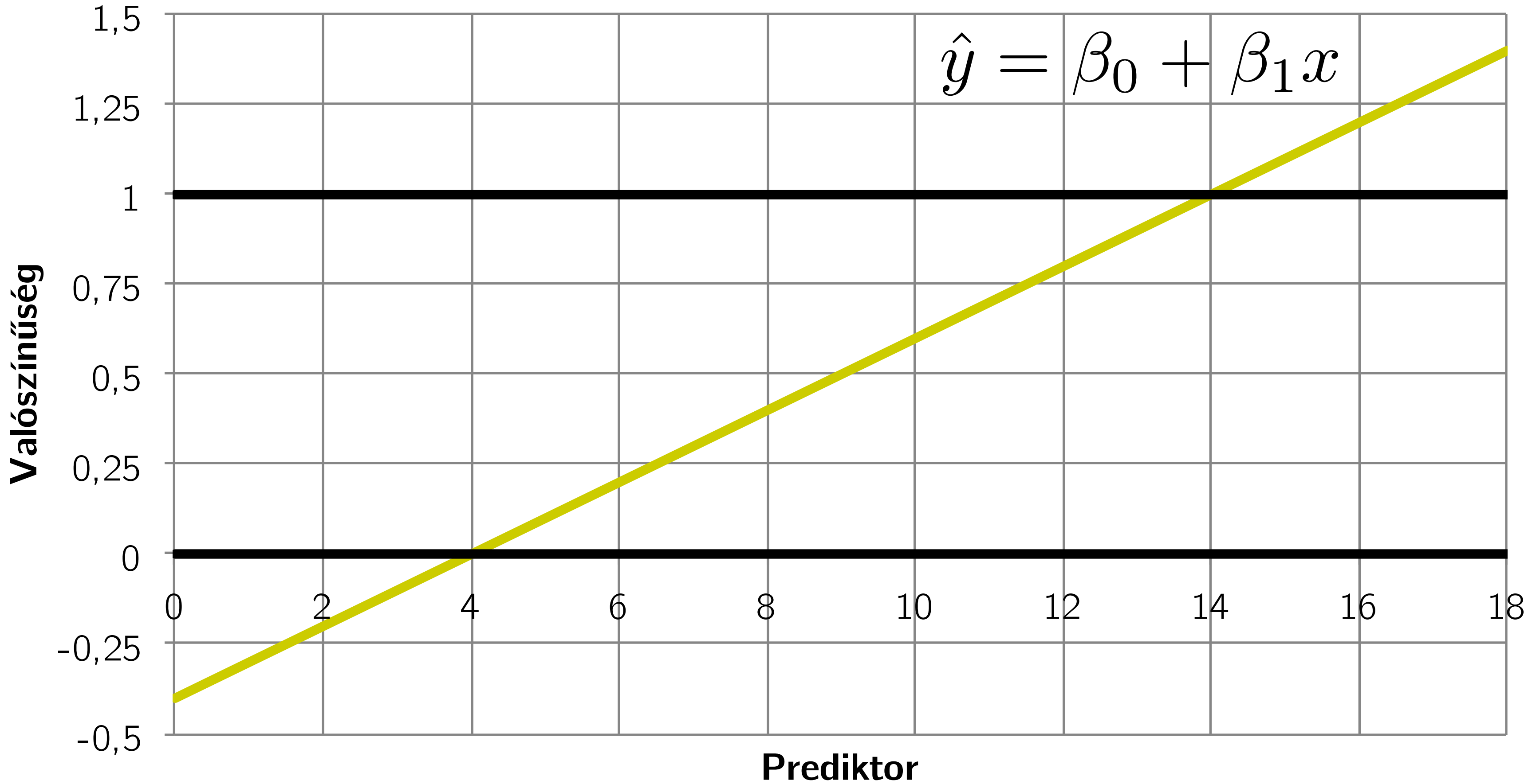
Bináris logit modell (logisztikus regresszió)

- Bináris (igen/nem) kimenetet modellez tetszőleges prediktorok függvényében.
- Becslést ad a két kimenet valószínűségére.
- A 20. század közepén az orvosi statisztikában terjedt el:
  Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association.*

Alkalmazások

- Mely ügyfeleknek nyújtson hitelt egy bank? (banki hitelelbírálás, credit scoring)
- Kik hagyják el mobilszolgáltatójukat a közeljövőben? (lemorzsolódás, churn analysis)
- Ki vesz meg várhatóan egy adott terméket? (marketing)
- Ki dolgozik várhatóan a következő évben? (nyugdíjrendszer modellezése)

Miért nem használható itt lineáris modell?



Logit modell

- Az y kimenet két kategóriáját 0 (R: ábécé-rendben az első) és 1 értékekkel kódoljuk.
- Egy pontszámot (logitot) rendelünk minden egyedhez, ami a prediktorok lineáris függvénye:

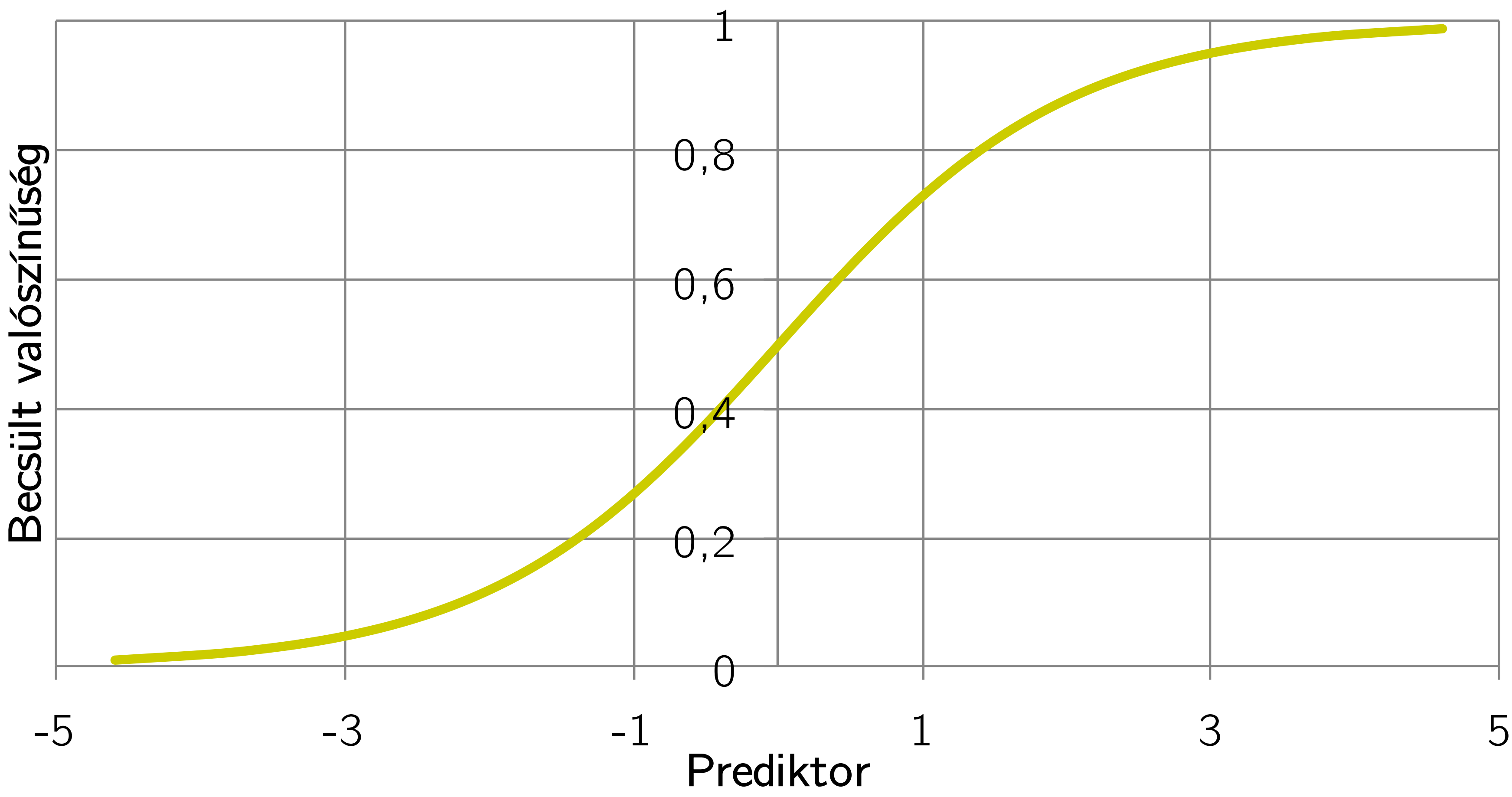
$$L = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- A logit elvben bármilyen értéket felvehet $-\infty$ és ∞ között. Ez még nem valószínűség!
- A jobb oldalon nincs hibatag! A bizonytalanság ott jelenik meg, hogy valószínűséget becsülünk.

Logit modell

- A logitot 0 és 1 közé transzformálva kapjuk a az $y=1$ (csőd) becsült valószínűségét:
$$\mathbb{P}(y=1) = \frac{1}{1+e^{-L}}$$
- A $\beta_0, \beta_1, \ldots, \beta_p$ együtthatókat maximum likelihood módszerrel becsüljük (lásd később).
- Ha egy prediktor együtthatója pozitív, akkor növelésének hatására nő az $y=1$ (csőd) becsült valószínűsége (*ceteris paribus!*).

Logisztikus függvény: a prediktorok hatása



Együtthatók maximum likelihood becslése

- Logitok és becsült valószínűségek:
$$L_i(\boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$
$$\mathbb{P}(y_i = 1) = \frac{1}{1 + e^{-L_i(\boldsymbol{\beta})}}$$
$$\mathbb{P}(y_i = 0) = 1 - P(y_i = 1)$$
- A likelihood annak a valószínűsége, hogy a kimenetek a megfigyelttel azonosan alakulnak:
$$L(\boldsymbol{\beta}) = \prod_{y_i=1} \mathbb{P}(y_i = 1) \prod_{y_i=0} \mathbb{P}(y_i = 0) \to \max.$$

Együtthatók maximum likelihood becslése

- A gyakorlatban a szoftver a likelihood függvény logaritmusát numerikusan maximalizálja (a megoldásra nincs képlet):

$$\ln L(\boldsymbol{\beta}) = \sum_{y_i=1} \ln \mathbb{P}(y_i = 1) + $$

$$+ \sum_{y_i=0} \ln \mathbb{P}(y_i = 0) \to \max .$$

Logit modellek típusai

- **Bináris** (*glm* függvény): a kimenetnek két kategóriája van.
- **Multinomiális** (*nnet* package, *multinom* függvény): a kimenet nominális, kettőnél több kategóriával.
- **Ordinális** (*MASS* package, *polr* függvény): a kimenet ordinális, kettőnél több kategóriával.
- Itt csak a bináris logit modellt tanuljuk!

Logit modellezés lépései

5. Értelmezés
4. Diagnosztika
3. Adatelőkészítés
2. Adatgyűjtés
1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: https://flaticon.com

Lineáris regressziós modellezés lépései

5. Értelmezés
4. Diagnosztika
3. Adatelőkészítés
2. Adatgyűjtés
1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: https://flaticon.com

Lineáris regressziós modellezés lépései

5. Értelmezés
4. Diagnosztika
3. Adatelőkészítés
2. Adatgyűjtés
1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: https://flaticon.com

Lineáris regressziós modellezés lépései

5. Értelmezés
4. Diagnosztika
3. Adatelőkészítés
2. Adatgyűjtés
1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: https://flaticon.com

Adatelőkészítés

- Adatok beszerzése, tisztítása, transzformációja, pótlása, stb.
- Pénzben kifejezett, *pozitív* (!) változókat (például ár, árfolyam, jövedelem, munkabér, vagyon stb.) logaritmikusan szokás transzformálni (így relatív, százalékos változásokat értelmezünk).
- Kategorikus prediktorokból dummy változókat kell képezni (R-ben automatikus).

Mintaméret

- Statisztikai hüvelykujj-szabály: legyen legalább 5-ször, de inkább 10-szer annyi megfigyelés, mint becsült paraméter.
- Különben a becslések nagyon bizonytalanok lesznek.
- Például ne építsünk 10 prediktorral modellt 27 EU-tagra!

Logit modellezés lépései

5. Értelmezés
4. Diagnosztika
3. Adatelőkészítés
2. Adatgyűjtés
1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: https://flaticon.com

Problémák a logit modellben

a. Kilógó értékek
b. Multikollinearitás
~~c. Hibatagok nemnormalitása~~
~~d. Heteroszkedaszticitás~~
e. Nemlinearitás
f. Felesleges prediktorok

Nincsenek hibatagok, ezért ez a két lineáris regressziós probléma nem merül fel a logit modellben!

Problémák a logit modellben

a. Kilógó értékek
b. Multikollinearitás
c. Nemlinearitás
d. Felesleges prediktorok

Diagnosztika (összefoglaló)

Jelenség	Miért baj?	Diagnózis	Megoldás
a. Kilógó értékek	Torzított modell	Stud. hibatagok	Megfigyelések elhagyása
b. Multikolli-nearitás	Bizonytalan együtthatók	VIF	Prediktorok elhagyása
c. Nemlinearitás	Torzított modell	RESET teszt, CR ábrák	Prediktorok transzformálása
d. Felesleges prediktorok	Nehéz értelmezés	z és χ^2 tesztek	Prediktorok elhagyása

Ugyanúgy ellenőrizhetjük és kezelhetjük a problémákat, mint a lineáris regresszióban (azzal a különbséggel, hogy t- és F-tesztek helyett z és χ^2–tesztek vannak).

Logit modellezés lépései

5. Értelmezés
4. Diagnosztika
3. Adatelőkészítés
2. Adatgyűjtés
1. Problémafelvetés

Forrás: saját szerkesztés, ikonok forrása: https://flaticon.com

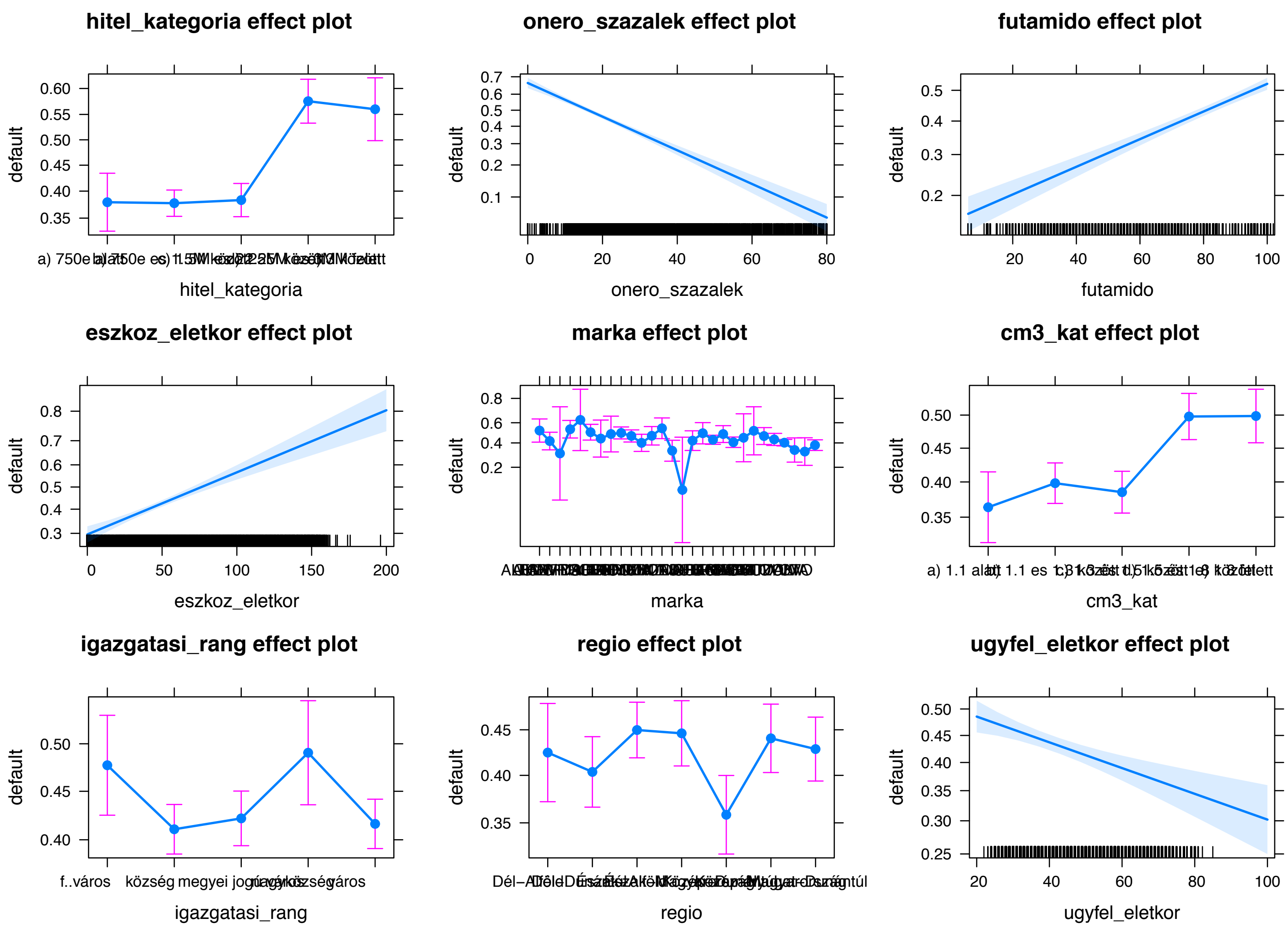
Értelmezés

a. Prediktorok parciális hatása
b. Prediktorok fontossága
c. Modell jósága, küszöbérték kalibrálása

a. Mi a prediktorok parciális hatása?

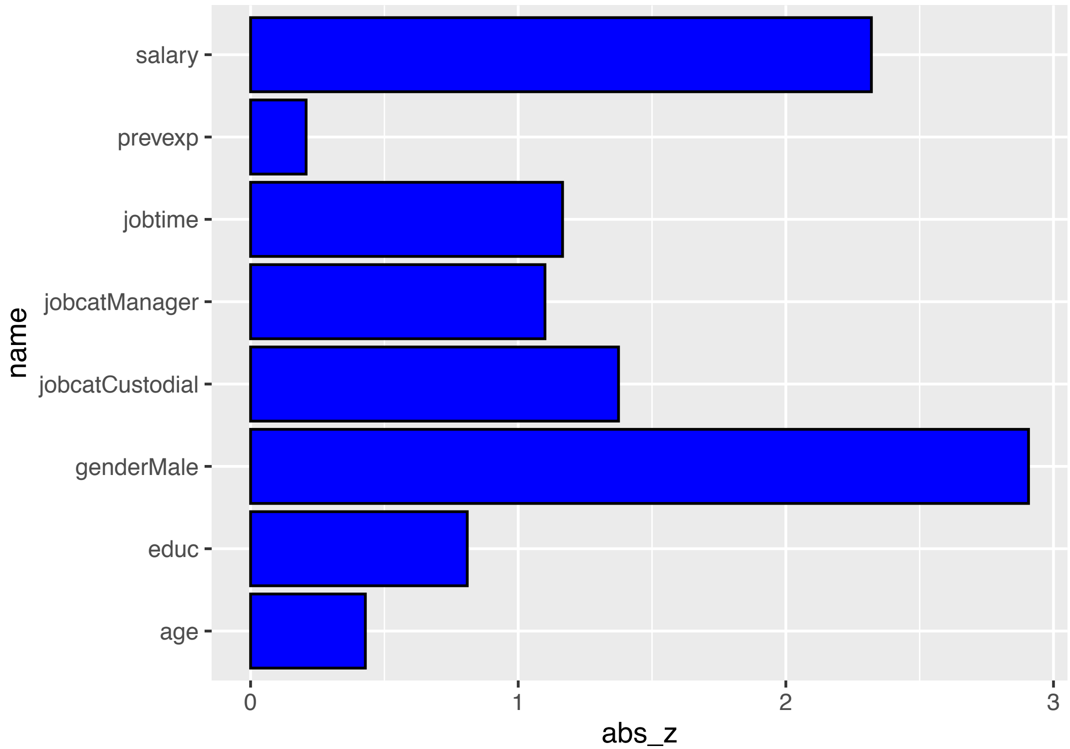
β_0	β_j (x_j numerikus)	β_j (x_j dummy)
Ha minden $x_j = 0$, akkor…	Ha x_j cet.par. egységnyivel nő, akkor…	Az adott kategóriában cet.par…
…a logit értéke β_0.	…az odds e^{β_j}-szeresére változik.	…az odds e^{β_j}-szerese a referencia-kategóriabeli odds-nak.

a. Mi a prediktorok parciális hatása?



b. Melyik prediktorok a legfontosabbak?

- A z-statisztikák abszolút értékei alapján rangsorolható a prediktorok fontossága.



c. Modell jósága és küszöbérték

- Ha a becsült valószínűség meghaladja a küszöbvalószínűséget (alapesetben $\frac{1}{2}$), a modell a kimenetet 1-nek, különben 0-nak becsüli.
- Osztályozó tábla:

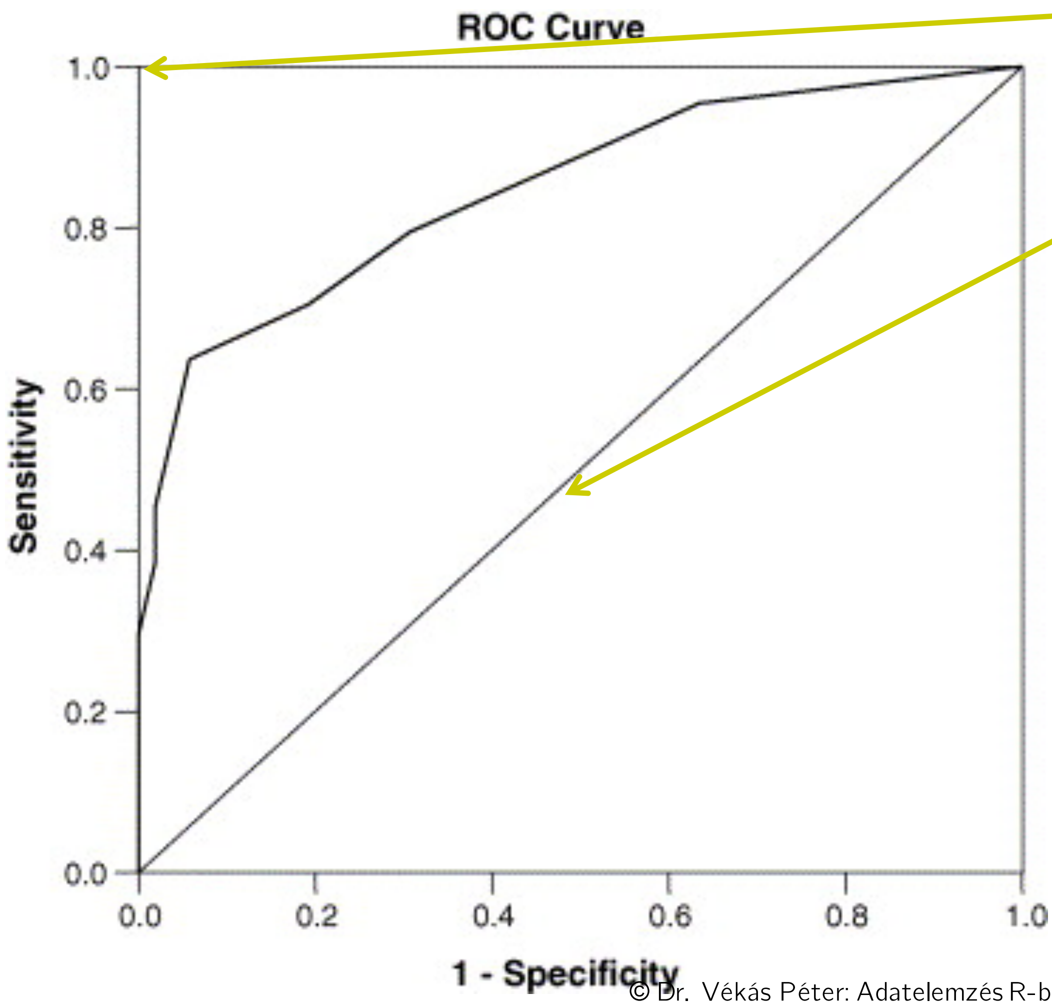
Tényleges \ Becsült	0	1
0	✓	Elsőfajú hiba
1	Másodfajú hiba	✓

- Találati arány: főátlóbeli elemek összege / n

ROC görbe értelmezése
(Receiver Operating Characteristic)

- Tengelyek:
  - x : tévesen besorolt megfigyelések (elsőfajú hiba) aránya a 0-s kategórián belül.
  - y: 1 – tévesen besorolt megfigyelések aránya az 1-es kategórián belül (1 – másodfajú hiba).
- Minden küszöbvalószínűséghez egy-egy (x,y) pár tartozik. Ezeket tartalmazza a ROC görbe.
- A görbe a 45 fokos egyenes felett halad.

Illeszkedés minősítése a ROC görbe alapján



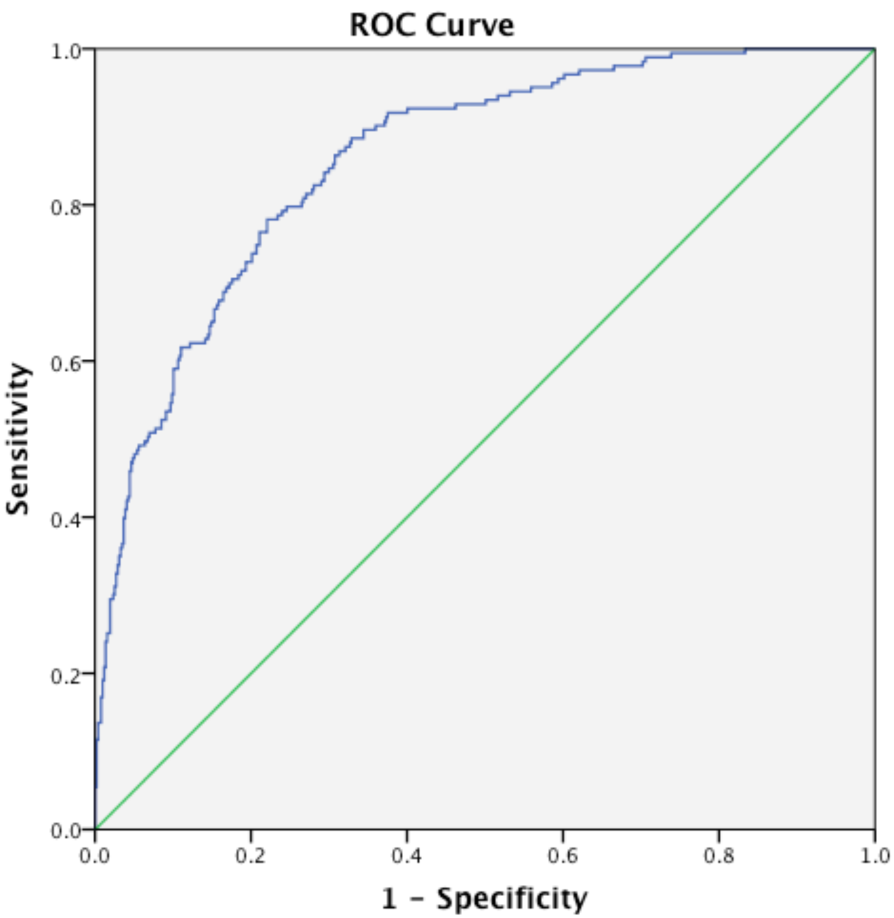
Hibátlan besorolás

Leggyengébb besorolás

Az illeszkedés mérőszámai:

- Görbe alatti terület (AUC, $0,5 \leq A \leq 1$)
- Gini-mutató ($Gini = 2AUC - 1$, $0 \leq G \leq 1$)

A modell illeszkedése



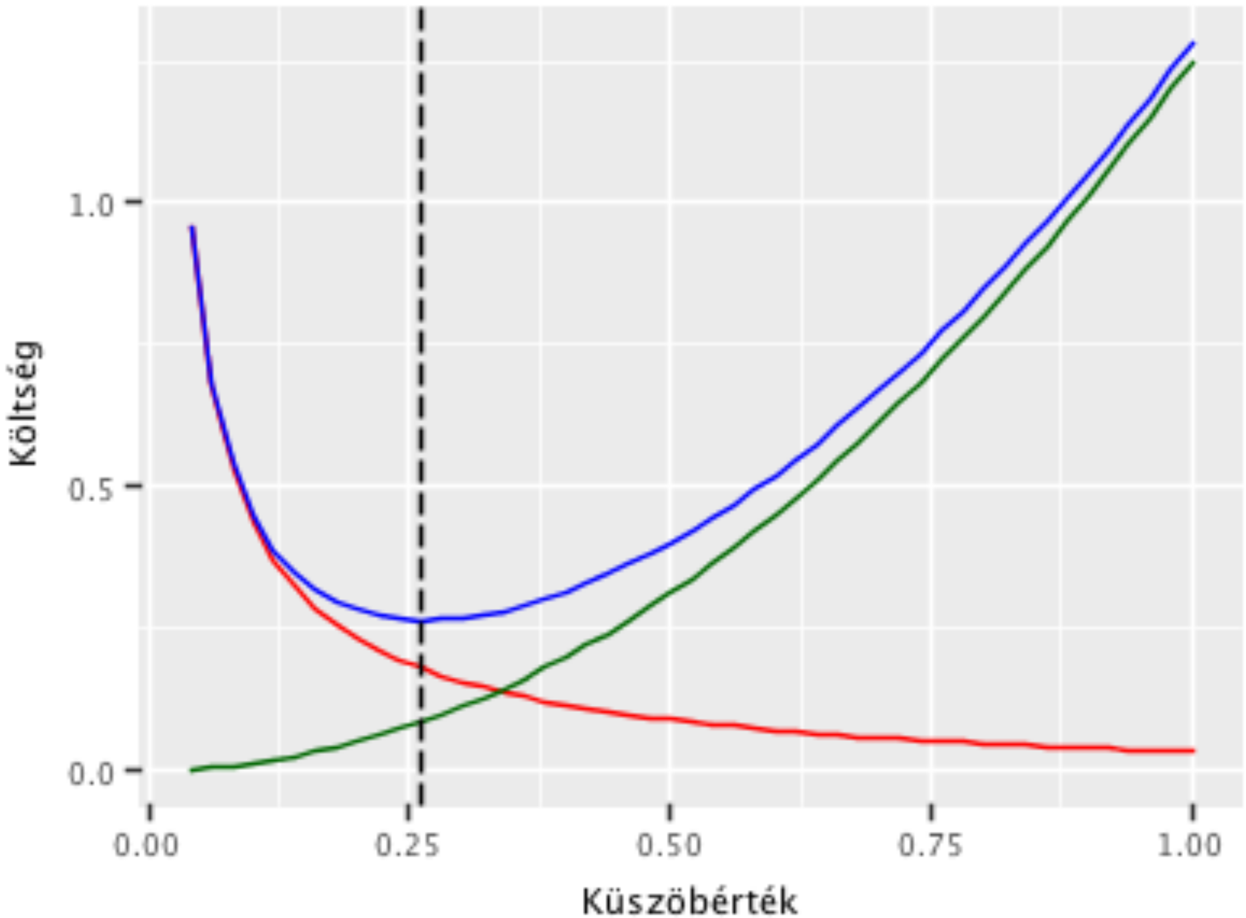
Area Under the Curve

Test Result Variable(s): Predicted probability

Area
0.858

AUC = 0,858,
Gini = 2 AUC − 1 = 0,716.

Költségfüggvény minimuma (elsőfajú, másodfajú, teljes)



Köszönöm a figyelmet!