Leíró és matematikai statisztika

Matematika alapszak, matematikai elemző specializáció

Zempléni András

Valószínűségelméleti és Statisztika Tanszék
Matematikai Intézet
Természettudományi Kar
Eötvös Loránd Tudományegyetem

Honlap: zempleni.elte.hu

E-mail: andras.zempleni@ttk.elte.hu

Szoba: D 3-310

12. előadás

Logisztikus regresszió

- Gyakran kell valószínűséget becsülnünk/osztályoznunk (két osztályba)
  - Betegség kialakulásának valószínűsége
  - Hitelező csődbemenetelének valószínűsége
  - A vizsga sikeres letételének valószínűsége
- Itt a hagyományos lineáris modell nem célravezető (könnyen adódnak negatív vagy 1-nél nagyobb értékek)
- A leggyakrabban használt, logisztikus függvény:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + \cdots + b_p x_p)}}$$

  ahol b_i a becsülendő együtthatók
- Az ún. odds-hányados: $P(Y = 1)/P(Y = 0) = e^{b_0 + b_1 x_1 + \cdots + b_p x_p}$
- Paraméterbecslés: pl. maximum likelihood módszerrel (numerikus módszerekkel lehet megkapni)

További kérdések, lehetőségek

- Az extrapoláció (a megfigyelt adatok tartományán kívülre történő előrejelzés) egyáltalán nem megbízható!
- Ha nem találtunk jól közelítő egyszerű függvényt, alkalmazhatunk nemparaméteres közelítést is a feltételes várható értékre (ez egyáltalán nem használható extrapolációra):

$$\widehat{E(Y|X=x)} = \frac{\sum_{i=1}^{n} Y_i k((x-X_i)/h_n)}{\sum_{i=1}^{n} k((x-X_i)/h_n)}$$

  ahol k a magfüggvény, h_n az ablakszélesség
- A Parzen-Rosenblatt tétel feltételei mellett konzisztens becslést ad
- Elnevezés: Nadarajah-Watson módszer
- Általánosítható lokális polinomiális közelítésre (loess)
- Az ablakszélesség lényeges (nem könnyű a jó megválasztása)
  - Ha túl kicsi, az egyedi megfigyelések zajosságát követi le a közelítés
  - Ha túl nagy, túlságosan sima eredményt kapunk

A vegyes kapcsolat elemzése – szóráselemzés

- A lineáris modell egyik legfontosabb alkalmazása (faktorok különböző szintjeinek van-e hatása?)
- Motivációs példák:
  - Hatással van-e egy vállalatnál a (bruttó) fizetésekre az, hogy valaki nő-e, avagy férfi?
  - Különböző vetőmagokra megnézték a termésátlagot egy nagyobb földterület különböző részein. Vajon hatással van-e a vetőmag fajtája a termésátlagra?
  - Hatással van-e a valszám gyakorlati összpontszámra, hogy a hallgatónak ki a gyakorlatvezetője?
- a megfigyelések y_{ij} az i-edik "szinten" mért j-edik érték
- szórásfelbontás: Teljes négyzetösszeg: $\mathrm{SST} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ij} - \overline{y_{\bullet\bullet}})^2$,

  külső (csoportok közötti) négyzetösszeg: $\mathrm{SSK} = \sum_{j=1}^{p} n_j (\overline{y_{j\bullet}} - \overline{y_{\bullet\bullet}})^2$,

  belső (csop.on belüli) négyzetösszeg: $\mathrm{SSB} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ij} - \overline{y_{j\bullet}})^2$

Vegyes kapcsolat mérőszáma

Szórásnégyzet-hányados

$$H^2 = 1 - \frac{SSB}{SST} = \frac{SSK}{SST}$$

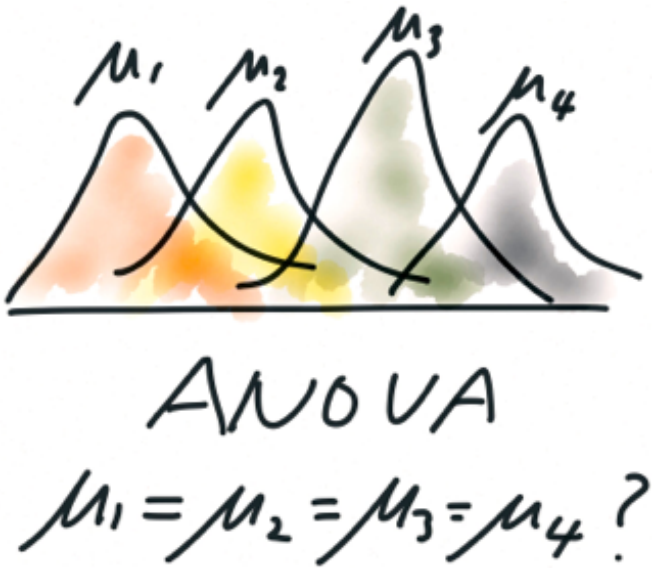
Megjegyzés: ez nem más, mint a regressziónál az R^2

Tulajdonságai:

- $H^2 = 0$ esetén a két ismérv között nincs a szórásfelbontással kimutatható kapcsolat, DE (!!) ezzel nem bizonyítottuk be, hogy függetlenek egymástól (analógia: korrelálatlanságból nem következik a függetlenség)
- $H^2 = 1$ esetén a két ismérv között függvényszerű kapcsolat van
- $0 < H^2 < 1$ esetén a két ismérv között sztochasztikus kapcsolat van
- erős a kapcsolat, ha H^2 közel van 1-hez és gyenge a kapcsolat, ha 0-hoz

Szóráselemzés (ANOVA)

- Elnevezései: szóráselemzés = variancia -analízis = ANOVA (analysis of variance)
- A szóráselemzési feladat fő kérdése: hatással van-e az eredményváltozó értékére, hogy a faktor melyik szintjén vagyunk? Jelölje b_i az i-edik szinten a várható értéket



$H_0 : b_1 = b_2 = \ldots = b_p$
H_1 : nem igaz H_0

ANOVA táblázat:

Szóródás forrása	Szabadság-fok	Négyzet-összegek	Tapasztalati szórásnégyzetek	
Külső	$p-1$	SSK	$MSK = \frac{SSK}{p-1}$	$F = \frac{\frac{SSK}{p-1}}{\frac{SSB}{n-p}}$
Belső	$n-p$	SSB	$MSB = \frac{SSB}{n-p}$	
Teljes	$n-1$	SST	–	

Szóráselemzés (ANOVA)

- F a H_0 esetén $(p-1, n-p)$ szabadságfokú F eloszlású. Ebből a kritikus tartomány: $F > f_{p-1,n-p,1-\alpha}$
- $\frac{\overline{y}_{i\bullet}-b_i}{\sqrt{MSB}}\sqrt{n_i} \sim t_{n-p}$ és $\frac{\overline{y}_{i\bullet}-\overline{y}_{j\bullet}-(b_i-b_j)}{\sqrt{MSB}}\sqrt{\frac{n_i n_j}{n_i+n_j}} \sim t_{n-p}$

Ezek alapján konfidenciaintervallumokat lehet készíteni a b_i értékekre és a $b_i - b_j$ különbségekre:

- Konfidenciaintervallumok:
  - b_i-re: $\overline{y}_{i\bullet} \pm t_{n-p;\alpha/2}\sqrt{\frac{MSB}{n_i}}$
  - $b_i - b_j$-re: $\overline{y}_{i\bullet} - \overline{y}_{j\bullet} \pm t_{n-p;\alpha/2}\sqrt{MSB}\sqrt{\frac{n_i+n_j}{n_i n_j}}$

Feladatok

E**50.)** A következő táblázatok a 2016/2017-es őszi félév Valószínűségszámítás c. tárgy hallgatóinak megoszlását mutatják aszerint, hogy a hallgató milyen szakos és a vizsgán hányast szerzett (csak azok szerepelnek, akik legalább 1-szer próbálkoztak).

a.)

Szakirány	1-es	2-es	3-as	4-es	5-ös	Összesen
Elemző	16	13	10	0	2	41
Infó A	4	3	7	1	7	22
Összesen	20	16	17	1	9	63

b.)

Szakirány	1-es	Legalább 2-es	Összesen
Elemző	16	25	41
Infó A	4	18	22
Összesen	20	43	63

Vizsgáljuk meg alkalmas mutatószámmal, hogy a megszerzett érdemjegyre hatással volt-e az, hogy a hallgató milyen szakra jár!

Feladatok

E51.) A következő táblázat az előző félévben tanár szakos BSc-s hallgatóknak tartott 4 bevezető valszám gyakorlat év végi, 100-ra skálázott végső pontszámait tartalmazza:

Gyakvezér	Pontszámok											
2*V.	98	87	102	92	52	46	95	60	81	55	60	94
	81	58	80	93	70	66	49	94	50	88	74	
2*G.	77	46	54	57	50	45	39	63	26	107	75	
	66	52	109	91	35	65						
2*Á.	86	94	54	61	42	59	88	81	81	80	102	72
	88	96	58	90	110	58	80	90	84	80	94	
2*L.	66	60	72	49	52	54	80	56	36	91	68	
	60	51	40	38	54	62						

a.) Vizsgáljuk meg, az év végi pontszám függött-e attól, hogy a hallgató melyik csoportba járt! Hány %-ban magyarázta a pontszámok változékonyságát az, hogy a hallgatók melyik csoportba jártak?

b.) Adjunk intervallumbecslést az egyes csoportok várható pontszámára!

c.) Állíthatjuk-e, hogy V. és Á. csoportjának átlagpontszámai (statisztikailag) egyenlők?

Kétszempontos szóráselemzés

- Gyakori eset, hogy nemcsak egy faktor hatását vizsgáljuk
  - Két gyógyszer együttesen hogyan hat?
  - Műtrágya és vetőmag-típus
  - Csoport és nem
- Az egyes faktorok külön-külön hatása mellett lényeges új kérdés a kölcsönhatás. Először ezt célszerű vizsgálni
- A modell: $Y_{ijm} = b_{ij} + \varepsilon_{ijm}$, ahol $i = 1, \ldots, p$ az első faktor, $j = 1, \ldots, r$ a második faktor szintjeit jelöli. $m = 1, \ldots, n$ az ismétlések száma. ε_{ijm} független, azonos eloszlású, 0 várható értékkel. Ez is lineáris modell.
- A szórásfelbontásnál új tag a kölcsönhatáshoz tartozó

$$SSI = m \sum_{i=1}^{p} \sum_{j=1}^{r} (\overline{y_{ij\bullet}} - \overline{y_{i\bullet\bullet}} - \overline{y_{\bullet j\bullet}} + \overline{y_{\bullet\bullet\bullet}})^2,$$

ennek szabadságfoka $(p-1)(r-1)$, az SSB szabadságfoka pedig $n - pr$. Ebből a szokásos módon kapható próba a kölcsönhatás hiányát leíró nullhipotézisre.