# Matematikai statisztika

## Informatika alapszak, "A" spec.

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

## 10. előadás

# A regressziós modell "felépítése"

Ha $p$ magyarázó változónk van, akkor $2^p$ modell közül kell a legjobbat megkeresni. Több módszer közül lehet választani:

- Nagyról kicsire (hátulról előre): először az összes magyarázó változót bevesszük, majd egyenként a legkevésbé szignifikánsat kivesszük egészen addig, míg mindegyik szignifikáns lesz
- Kicsiről nagyra (elölről hátulra): egyesével azt vesszük hozzá, amelyikkel a legjobban illeszkedő modellt kapjuk a következő lépésben.
  Vége: ha bármelyik, még a modellen kívüli magyarázó változót bevéve, már nem javul a modell illeszkedése.

Ha nominális skálán mért magyarázó változóink (is) vannak, akkor úgynevezett "dummy" változókat vezethetünk be, amik indikátorok, megfelelnek az egyes értékeknek ($i = 1, .., k - 1$), mert az utolsó már kizámolható abból, hogy minden megfigelésre ezek közül pontoan az egyik következik be

# Logisztikus regresszió

- Gyakran kell valószínűséget becsülnünk
  - 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:

$$P(Y=1) = \frac{1}{1+e^{-(b_0+b_1x_1+\cdots+b_px_p)}}$$

  ahol $b_i$ a becsülendő együtthatók
- az ún. odds-hányados: $P(Y=1)/P(Y=0) = e^{b_0+b_1x_1+\cdots+b_px_p}$
- Paraméterbecslés: pl. maximum likelihood módszerrel (numerikus módszerekkel lehet megkapni)

# 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) 2

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

# Sűrűségfüggvény becslése – magfüggvényes módszer (Parzen-Rosenblatt becslés)

$$f_n(x) = \frac{1}{n \cdot h_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \text{ ahol}$$

- $K : \mathbb{R} \to \mathbb{R}$ páros sűrűségfüggvény, neve: magfüggvény
- $h_n$ sávszélesség, rendszerint $h_n = n^c$, ahol $-1 < c < 0$ valós szám

A leggyakoribb magfüggvények sűrűségfüggvény becslésére:

| Magfüggvény neve | $K(x)$ |
|---|---|
| Gauss | $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ |
| téglalap (rectangular) | $\frac{1}{2} I(\lvert x \rvert \leq 1)$ |
| háromszög (triangular) | $(1 - \lvert x \rvert) \cdot I(\lvert x \rvert \leq 1)$ |
| Bartlett–Epanechnikov | $\frac{3}{4}(1 - x^2) \cdot I(\lvert x \rvert \leq 1)$ |
| cosinus | $\frac{\pi}{4} \cos\left(\frac{\pi}{2} x\right) \cdot I(\lvert x \rvert \leq 1)$ |

# Sűrűségfüggvény becslése: példák

Mi az "optimális" sávszélesség? Mi az, hogy "optimális"? $\rightsquigarrow$ amivel $f_n(x)$ "legjobban" közelíti a valódi sűrűségfüggvényt. Függ a becsülendő eloszlástól, így nincs univerzális válasz

E**10.)**

a.) Legyen a rendezett mintánk 1,2,5,6,12. Számoljuk ki a sűrűségfüggvény becslését, ha $h = 0,5$ és a téglalap-magfüggvényt alkalmazzuk!

b.) Mennyiben változik a becslés, ha a háromszög-magfüggvényt választjuk?

c.) Mennyi az R által alkalmazott default sávszélesség?

# Alkalmazás a regresszionál, további kérdések

- 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