Matematikai statisztika

Informatika alapszak, "B" 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

6. stat. előadás

Lineáris modell

- A modell: $\mathrm{y} = X\mathrm{b} + \varepsilon$
- $F := \mathrm{Im}X \quad \rightsquigarrow X$ képtere
- $r :=$rang(X), általánosan $r \leq p$, teljes rangú esetben $r = p$
- Paraméterbecslés: $\widehat{\mathrm{b}} = (X^T X)^{-1} X^T \mathrm{y}$
- Projekció az F altérre: $P_F = X(X^T X)^{-1} X^T$
- Becsült értékek: $\widehat{\mathrm{y}} := X\widehat{\mathrm{b}}$
- Reziduálisok: $\widehat{\varepsilon} = \mathrm{y} - \widehat{\mathrm{y}}$
- Reziduális négyzetösszeg: $\mathrm{RNÖ} := \mathrm{SSR}\|\widehat{\varepsilon}\|^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$
- Teljes négyzetösszeg: $\mathrm{NÖ}(= \mathrm{SS}) = \sum_{i=1}^{n} (y_i - \overline{y})^2$
- Determinációs együttható: $R^2 = 1 - \frac{\mathrm{RNÖ}}{\mathrm{NÖ}} = \frac{\mathrm{NÖ} - \mathrm{RNÖ}}{\mathrm{NÖ}} \quad \rightsquigarrow$ az eredményváltozó változékonyságának hány %-át magyarázza a regressziós modell
  Értéke 0 és 1 között lehet. Minél nagyobb, annál jobb.

Lineáris modell

- Korrigált determinációs együttható: $R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}\frac{SSR}{SS}$ $\rightsquigarrow$ egy lehetséges modellválasztási kritérium, minél nagyobb, annál jobb
- Akaike-féle információs kritérium: $AIC = 2(p+1) - 2\log \widehat{L}$, ahol $\widehat{L}$ a likelihood-függvény értéke akkor, ha az ML-becslést használjuk (normális eloszlású hibáknál ez megegyezik a legkisebb négyzetes becsléssel)
  Ez is egy lehetséges modellválasztási kritérium, minél kisebb, annál jobb.

Hipotézisvizsgálat a lineáris modellben

- t-próba az egyes együtthatókra (feltételezzük a hibák normális eloszlását): $H_0 : b = 0$, $H_1 : b \neq 0$
- A próbastatisztika: $t = \frac{\hat{b}}{D(\hat{b})}$, ez $n-1$ szabadságfokú t-eloszlású, ha igaz a H_0.
- Az egy magyarázó változós esetben ($y \sim a + bx$) $D^2(b) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$ és $D^2(a) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right]$.
- A reziduálisok: $r_i = y_i - \hat{y}_i$, ebből becsülhető a hiba szórásnégyzet: $\hat{\sigma}^2 = \sum_{i=1}^{n} r_i^2/(n-p)$
- A szokott módon konfidencia intervallum is konstruálható a becslésekhez

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.

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:

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

  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)

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
- Elnevezés: Nadarajah-Watson módszer
- 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

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

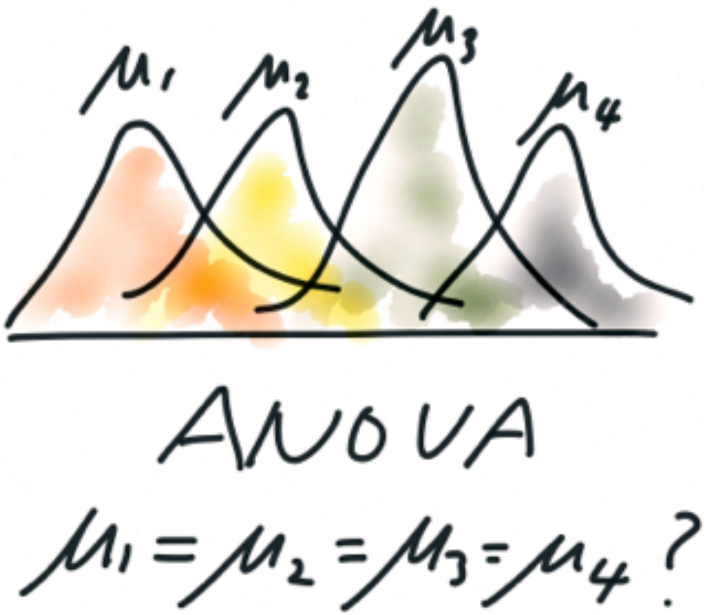
Megjegyzés: ez nem más, mint a regressziónál az R^2

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	–	