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

Regresszióelemzés, lineáris modell – motiváció

- Gyakorlati szempontból a félév egyik legfontosabb témája!
- Az 1. órai kérdőíves felmérés alapján mennyire magyarázható jól
  - a hallgatók testmagassága a súlyuk segítségével?
  - a hallgatók testmagassága a súlyuk és a cipőméretük segítségével?
  - a hallgatók statisztika érdemjegye a testmagasságuk segítségével?
- Egy részvény holnapi árfolyamát hogyan jelezzük előre a tegnapi, tegnapelőtti, stb. árfolyamadatok segítségével?
- Egy gazda földvásárlási dilemmája – egy bizonyos földterületen a várható termésátlag mennyire jelezhető előre a földterület fontosabb jellemzői alapján (a talaj kémhatása, a $CaCO_3$ megjelenési mélysége, a humusztartalom, topográfiai helyzet)?
- Meg lehet-e becsülni annak az esélyét, hogy valaki élete során megbetegszik tüdőrákban? Hogyan modellezzük ezt? Például: megbetegedés esélye $\longleftarrow$ dohányzik-e, hány éven át dohányzott élete során, van-e tüdőrákos a közeli rokonságban, van-e egyéb tüdőbetegsége, poros/füstös helyen dolgozik-e?

Regresszióelemzés – bevezetés

Legyenek $Y, X_1, \ldots, X_p$ véges szórású valószínűségi változók, amik egy véletlen jelenség egy-egy jellemzői.

A regresszióelemzés célja: a bennünket különösen érdeklő Y valószínűségi változó "minél jobb" közelítése az $X_1, \ldots, X_p$ valószínűségi változók segítségével.

Y elnevezései: eredményváltozó, függő változó, endogén változó
X_i-k elnevezései: magyarázó változók, független változók, exogén változók

Általában megfigyeléseink vannak, amik az $(Y, X_1, \ldots, X_p)^T$ valószínűségi vektorváltozó realizációinak tekinthetők:
$(y_i, x_{i,1}, \ldots, x_{i,p})^T \quad i = 1, 2, \ldots, n \qquad$ általában $n >> p$
Feltehetjük, hogy az y_i megfigyelések rendszerint mérési eredmények, amik sajnos pontatlanok. A mérési hibát ε_i-vel fogjuk jelölni, amiről természetes feltétel, hogy legyen 0 várható értékű és véges σ szórású valószínűségi változó.

Regresszióelemzés: elméleti eredmények

Legyenek Y, X, $X_1, \ldots, X_p$ véges szórású valószínűségi változók, $c, a, b_1, \ldots, b_p$ valós számok.
Jelölje $\mathbf{X} = (X_1, \ldots, X_p)^T$, $\mathbf{b} = (b_1, \ldots, b_p)^T$ vektorokat.

	Feladat	Megoldás
a.)	$\min_c E(Y-c)^2$	$\hat{c} = EY$
b.)	$\min_{f:\mathbb{R}\to\mathbb{R} \text{ mérhető fv.}} E(Y-f(X))^2$	$\widehat{f}(X) = E(Y\|X)$
c.)	$\min_{a,b} E(Y-(a+bX))^2$	$\hat{b} = \frac{cov(X,Y)}{D^2X}$, $\hat{a} = EY - \hat{b}EX$
d.)	$\min_{f:\mathbb{R}^p\to\mathbb{R} \text{ mérhető fv.}} E(Y-f(X_1,\ldots,X_p))^2$	$\widehat{f}(X_1,\ldots,X_p) = E(Y\|X_1,\ldots,X_p)$
e.)	$\min_{a,b_1,\ldots,b_p} E\left(Y-(a+\sum_{i=1}^{p} b_iX_i)\right)^2$ [Többváltozós lineáris regresszió]	$\hat{\mathbf{b}} = (cov(\mathbf{X},\mathbf{X}))^{-1} cov(\mathbf{X}, Y)$ $\hat{a} = EY - \sum_{i=1}^{p} \hat{b}_i EX_i$

$E(Y|X)$: feltételes várható érték

Lineáris modell

- A modell: $\mathbf{y} = X\mathbf{b} + \varepsilon$
- $F := \mathrm{Im}X \quad \rightsquigarrow X$ képtere
- $r :=$rang(X), általánosan $r \leq p$, teljes rangú esetben $r = p$
- Paraméterbecslés: $\widehat{\mathbf{b}} = (X^T X)^{-1} X^T \mathbf{y}$
- Projekció az F altérre: $P_F = X(X^T X)^{-1} X^T$
- Becsült értékek: $\widehat{\mathbf{y}} := X\widehat{\mathbf{b}}$
- Reziduálisok: $\widehat{\varepsilon} = \mathbf{y} - \widehat{\mathbf{y}}$
- 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 ($y \sim a + bx$) $D^2(\hat{b}) = \frac{\sigma^2}{\sum(x_i - \overline{x})^2}$ és $D^2(\hat{a}) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right]$.
- 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.