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

11. előadás

Regresszióelemzés

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

$E(Y|X)$: feltételes várható érték

Eszköz: feltételes várható érték

- Adottak X, Y valószínűségi változók, amelyek között van összefüggés
- Szeretnénk Y-ból kinyerni minden X-re vonatkozó információt
- Az Y-ban lévő információt Y függvényei jelentik $\Leftrightarrow$ azt a $g(Y)$ valószínűségi változót szeretnénk meghatározni, amely leginkább hasonlít X-re, legközelebb van X-hez.
- A feladat általános megoldása egy alkalmas térben: nézzük a $g(Y)$-ok által kifeszített alteret és megkeressük az X merőleges vetületét erre az altérre $\Rightarrow$ ez lesz az X-hez legközelebbi $g(Y)$ alakú valószínűségi változó.
- Ez a vetület X-nek Y szerinti feltételes várható értéke: $g(Y) = E(X|Y)$

Feltételes várható érték

Definíció: Ha X diszkrét valószínűségi változó és $P(B) > 0$, akkor X feltételes várható értéke a B feltétel mellett: $X \sim \begin{matrix} x_1, x_2, \dots \\ p_1, p_2, \dots \end{matrix}$

$$E(X|B) = \sum_{i=1}^{\infty} x_i \cdot P(X = x_i|B)$$

Valószínűségi változó szerinti feltételes várható érték X, Y diszkrét valószínűségi változó $X = \begin{matrix} x_1, x_2, \dots \\ p_1, p_2, \dots \end{matrix}$, $Y = \begin{matrix} y_1, y_2, \dots \\ q_1, q_2, \dots \end{matrix}$, $EX < \infty$
$Y = y_i$ esemény és $P(Y = y_i) > 0$, így $E(X|Y = y_i)$ definiált a fentiek szerint.
Definíció: $E(X|Y = y) = g(y)$ egy függvény az Y értékkészletén az $y_i \to E(X|Y = y_i)$ hozzárendelés szerint.

Valószínűségi változó szerinti feltételes várható érték

Definíció: Ha $g(y) = E(X|Y = y)$ az X feltételes várható értéke az $Y = y$ feltétel mellett, akkor a $g(Y) = E(X|Y)$ az X feltételes várható értéke az Y feltétel (vagy adott Y) mellett. $E(X|Y)$ valószínűségi változó.

Megjegyzés: Ha $Y(\omega) = y_i$, akkor $E(X|Y)(\omega) = g(y_i) = g(Y(\omega))$

Tulajdonságok:

- Lineáris : $E(c \cdot X_1 + X_2|Y) = c \cdot E(X_1|Y) + E(X_2|Y)$
- Ha $X \geq Z \Rightarrow E(X|Y) \geq E(Z|Y)$
- Ha X, Y függetlenek: $E(X|Y) = EX$
- Ha $X = h(Y)$, akkor $E(X|Y = y_j) = h(y_j)$ ezért $E(X|Y = y) = h(y) \Rightarrow E(X|Y) = h(Y) = X$
- $E(E(X|Y)|Y) = E(g(Y)|Y) = g(Y) = E(X|Y)$

A feltételes várható érték kiszámítása

Az abszolút folytonos eset: legyen X, Y együttes sfv.-e $f(x,y)$. X, ill. Y sfv.-e: $\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y = f_X(x)$, $\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x = f_Y(y)$

Definíció: A feltételes sűrűségfüggvény:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x}$$

Definíció: A feltételes eloszlásfüggvény: $F(x|y) = \int_{-\infty}^{x} f(t|y)\mathrm{d}t$

Definíció: A feltételes valószínűség y függvényében:

$$P(X \in A|\, Y = y) = \int_A f(t|y)\mathrm{d}t$$

Teljes valószínűség tétele

Állítás:

$$P(X \in A, Y \in B) = \int_A \int_B f(x|y) f_Y(y) \mathrm{d}y \mathrm{d}x =$$

$$\overset{\text{integrál csere}}{=} \int_B P(X \in A | Y = y) f_Y(y) \mathrm{d}y$$

Legyen $g(y) = \int_{-\infty}^{+\infty} x \cdot f(x|y) \mathrm{d}x$, akkor $g(Y) = E(X|Y)$

Következmény: Beírva $f(x|y)$ definícióját is kapjuk, hogy

$$g(y) = \int_{-\infty}^{+\infty} x \cdot \frac{f(x,y)}{f_Y(y)} \mathrm{d}x = \int_{-\infty}^{+\infty} \frac{x \cdot f(x,y)}{\int_{-\infty}^{+\infty} f(t,y) \mathrm{d}t} \mathrm{d}x$$

Példák

E47.) Legyen X és Y együttes sűrűségfüggvénye $h(x,y) = \exp(-y)$, ha $0 < x < y$, és 0 máshol. $E(X|Y) = ?$

E48.) Legyen X és Y együttes sűrűségfüggvénye

$$h(x,y) = \frac{12}{5}(x+y) \text{ ha } 0 < \frac{x}{2} \leq y \leq 1 - \frac{x}{2}$$

és 0 különben. $E(X|Y) = ?$

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.

Feladatok

E49.) Tekintsünk az alábbi regressziós modellekre lineáris modellként, és becsüljük meg a paramétereket! Jelölések: (y_i, x_i) a megfigyelések, ε_i a mérési hiba $(i = 1, \ldots, n)$, a becsülendő paraméterek pedig a, b, c.

a.) $y_i = a + bx_i + \varepsilon_i \quad \rightsquigarrow$ (egyszerű) kétváltozós regresszió
b.) $y_i = a + bx_i + cx_i^2 + \varepsilon_i \quad \rightsquigarrow$ négyzetes regresszió
c.) $y_i = a + b \sin x_i + c \cos x_i + \varepsilon_i \quad \rightsquigarrow$ harmonikus regresszió

Határozzuk meg a becsült paramétereket **R** segítségével és ábrázoljuk a megfigyeléseket az illesztett görbével együtt, ha a megfigyelések a következők:

y_i	-0,82	1,72	2,72	1,14	0,96	0,93	-0,08	0,29	3,38	3,36
x_i	3,92	2,63	1,68	2,57	2,61	2,78	3,81	2,89	0,28	0,94

Értékeljük az egyes modelleket önmagukban, és egymáshoz képest is!