Az Agrármérnöki MSc szak tananyagfejlesztése
TÁMOP-4.1.2-08/1/A-2009-0010 projekt

ÁILATGENETIKA

Debreceni Egyetem
Nyugat-magyarországi Egyetem
Pannon Egyetem

A projekt az Európai Unió támogatásával,
az Európai Szociális Alap
társfinanszírozásával valósul meg.

Lineáris algebra
és
lineáris modellek

Lineáris algebra

Mátrix: általában számokból (változókból) alkotott sorokban és oszlopokban rendezett táblázat.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m1} & \cdots & a_{mn} \end{bmatrix} = [a_{ik}]$$

Minden mátrix **n sorból** és **m oszlopból** áll, így a mátrixokat **n x m dimenziósnak (méretűnek)** szokás nevezni.

A egyetlen sorból álló mátrixot ***sormátrixnak***, az egyetlen oszlopból álló mátrixot ***oszlopmátrixnak*** nevezzük.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \end{bmatrix} \qquad B = \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix}$$

Az egyetlen sorból és egyetlen oszlopból álló mátrix egyetlen szám (***skalár***).

Mátrixok darabolása

A kiindulási mátrixot több részmátrixra bontjuk

$$C = \begin{bmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{bmatrix} = \left[\begin{array}{c|cc} 3 & 1 & 2 \\ \hline 2 & 5 & 4 \\ 1 & 1 & 2 \end{array}\right] = \begin{bmatrix} a & b \\ d & B \end{bmatrix}$$

$$a = \begin{bmatrix} 3 \end{bmatrix} \qquad b = \begin{bmatrix} 1 & 2 \end{bmatrix} \qquad d = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \qquad B = \begin{bmatrix} 5 & 4 \\ 1 & 2 \end{bmatrix}$$

Mátrixműveletek

Összeadás

$$A+B=\left[a_{ik}\right]_{mn}+\left[b_{ik}\right]_{mn}=\left[a_{ik}+b_{ik}\right]_{mn}=C$$

Kivonás

$$A-B=\left[a_{ik}\right]_{mn}-\left[b_{ik}\right]_{mn}=\left[a_{ik}-b_{ik}\right]_{mn}=D$$

$$\mathrm{A}=\begin{bmatrix}3 & 0\\ 1 & 2\end{bmatrix} \qquad \mathrm{B}=\begin{bmatrix}1 & 2\\ 2 & 1\end{bmatrix}$$

$$\mathrm{C}=\mathrm{A}+\mathrm{B}=\begin{bmatrix}4 & 2\\ 3 & 3\end{bmatrix} \qquad \mathrm{D}=\mathrm{A}-\mathrm{B}=\begin{bmatrix}2 & -2\\ -1 & 1\end{bmatrix}$$

Feltétele: A két mátrix dimenzióinak meg kell egyeznie.

Szorzás

Szorzás skalárral

$$C = k * A = [k * a_{ij}]$$

$$C = -2 * \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ -6 & -2 \end{bmatrix}$$

Két mátrix szorzása

$$C_{n\times k} = A_{n\times m} * B_{m\times k} \quad C_{ij} = a_{i1} * b_{1j} + a_{i2} * b_{2j} + \cdots + a_{ip} * b_{pj} = \sum_{k=1}^{p} a_{ik} * b_{kj}$$

Fontos! $A * B \neq B * A$

$$C = \begin{bmatrix} 2 \\ 5 \end{bmatrix} * [3 \quad 4] = \begin{bmatrix} 6 & 8 \\ 15 & 20 \end{bmatrix}$$

Mátrixműveletek szabályai

$$A * (B + C) = A * B + A * C$$

$$(A + B) * C = A * C + B * C$$

$$A + (B * C) = (A * B) * C$$

$$(A * B)^T = B^T * A^T$$

Transzponálás (AT v. A’)

Egy A mátrixból a sorok és oszlopok felcserélésével képzett mátrixot az A mátrix ***transzponáltjának*** nevezzük.

$$A = \begin{bmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{bmatrix} \qquad A^T = \begin{bmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{bmatrix}^T = \begin{bmatrix} 3 & 2 & 1 \\ 1 & 5 & 1 \\ 2 & 4 & 2 \end{bmatrix}$$

$$(A * B * C)^T = C^T * B^T * A^T$$

Egységmátrix (E)

$$A * E = A \qquad E * A = A$$

$$E = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$E_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$$

Invertálás (A^{-1})

$$A^{-1} * A = E = A * A^{-1}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad A^{-1} = \frac{1}{a*d - b*c} * \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Ha det(A)<>0, az inverz létezik, és az A mátrix nem-szinguláris. Ha det(A)=0, az A mátrix szinguláris, és nem létezik egyetlen inverze (általánosított inverze viszont igen).

$$(A * B)^{-1} = B^{-1} * A^{-1}$$

Invertálás (A^{-1})

$$A = \begin{bmatrix} 3 & 5 \\ -2 & 4 \end{bmatrix}$$

$$A^{-1} = \frac{1}{3*4-5*(-2)} * \begin{bmatrix} 4 & -5 \\ 2 & 3 \end{bmatrix} = \frac{1}{22} * \begin{bmatrix} 4 & -5 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 4/22 & -5/22 \\ 2/22 & 3/22 \end{bmatrix}$$

Inverz mátrix felhasználása – egyenletrendszerek megoldása

Kiszámítandó ismeretlenek oszlopvektora

Ismert együtthatók mátrixa

Ismert együtthatók oszlopvektora

$$A * x = c$$

$$A^{-1} * A * x = A^{-1} * c$$

$$x = A^{-1} * c$$

Inverz mátrix felhasználása – egyenletrendszerek megoldása - példa

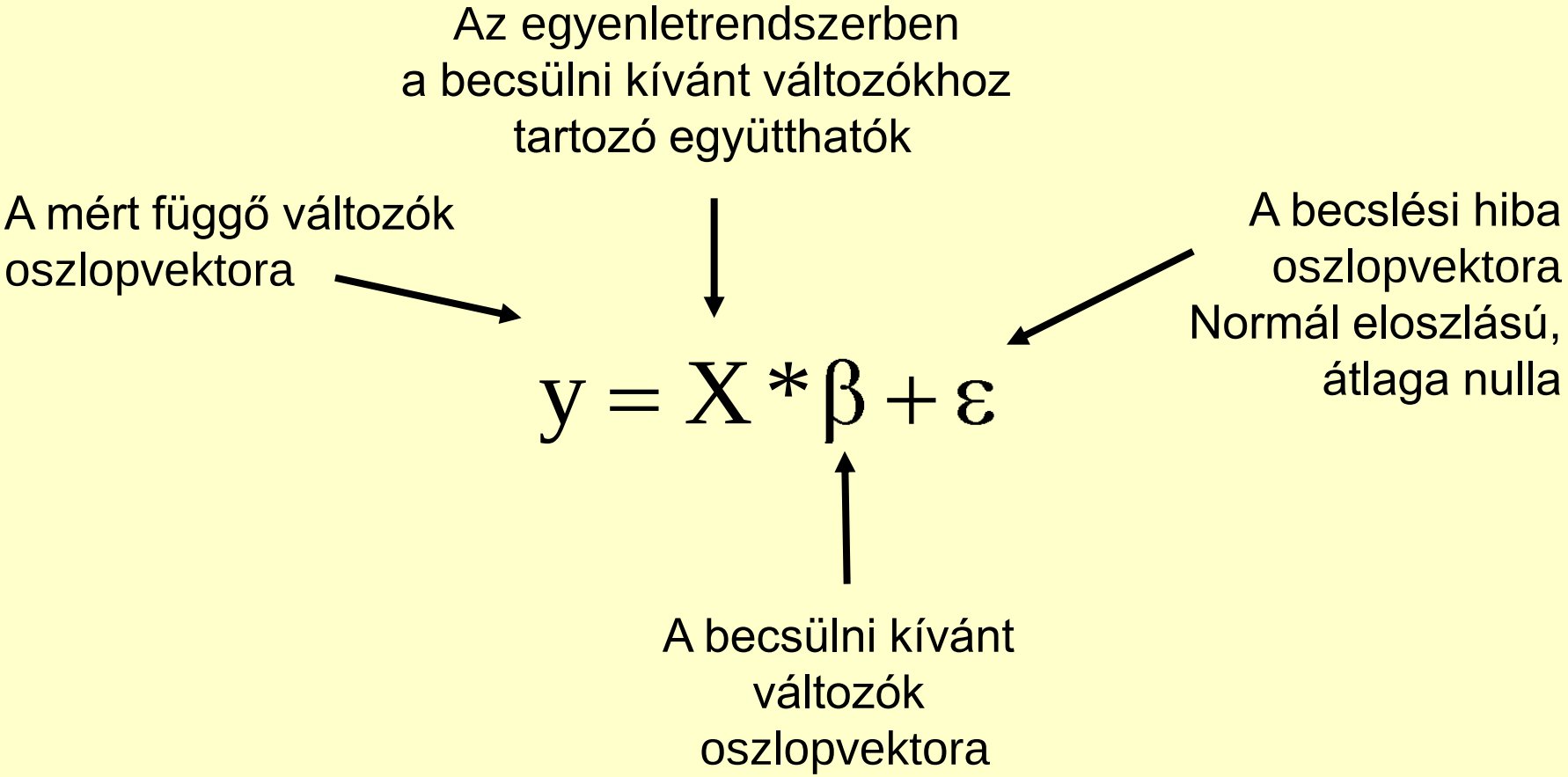
$$x_1 + 3x_2 + 3x_3 = 1$$
$$x_1 + 3x_2 + 4x_3 = 2$$
$$x_1 + 4x_2 + 3x_3 = 1$$

$$A = \begin{bmatrix} 1 & 3 & 3 \\ 1 & 3 & 4 \\ 1 & 4 & 3 \end{bmatrix} \quad c = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad A^{-1} = \begin{bmatrix} 7 & -3 & -3 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}$$

$$x = A_{-1} * c = \begin{bmatrix} 7 & -3 & -3 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} * \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

Lineáris modellek

Az általánosított lineáris modell alakja:



Lineáris modellek

A függő „y” változó lineáris függvénnyel történő leképezése több független (vagy becslő) változóval.

A többváltozós regressziós modell a legegyszerűbb lineáris modell.

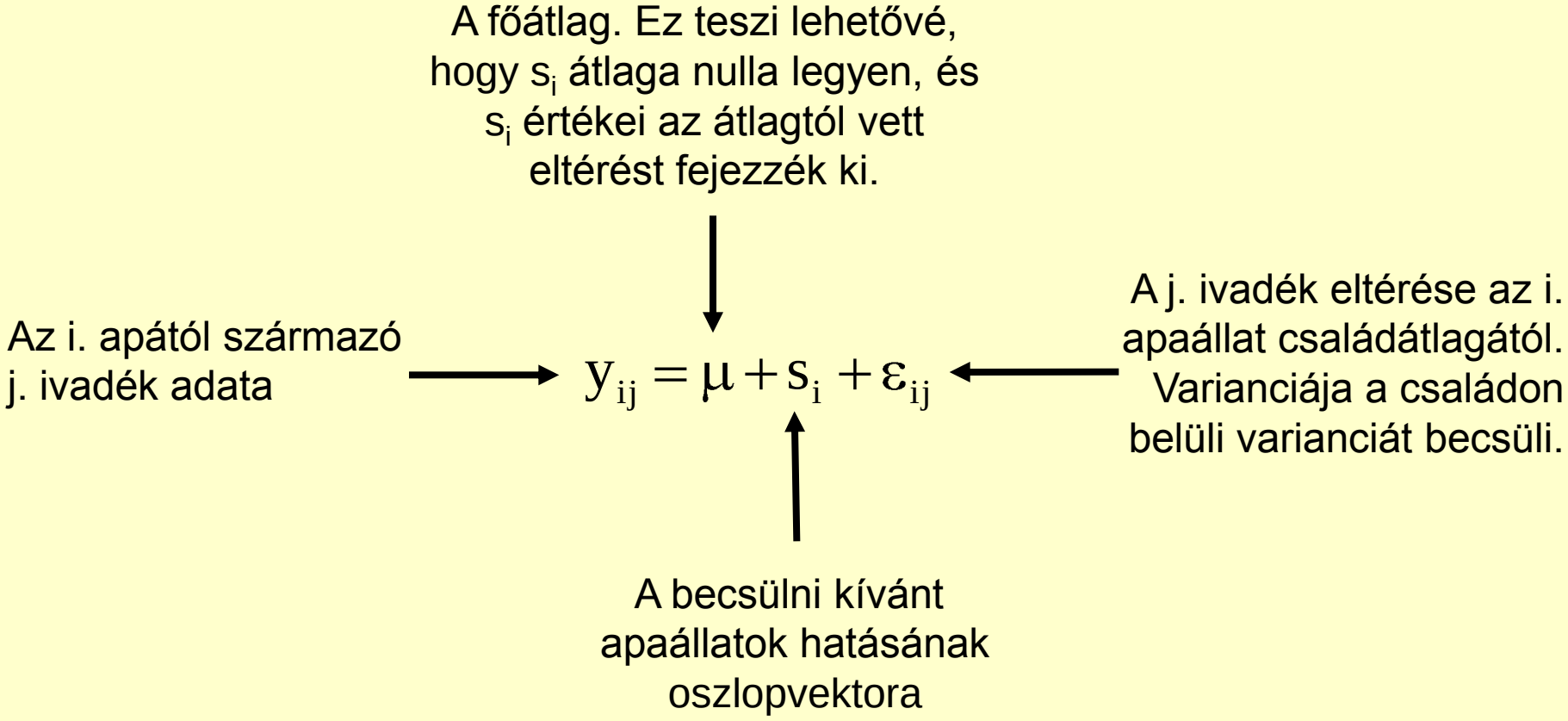
$$y = \mu + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

A β_i regressziós együtthatók jelentése: Az x_i egy egységnyi változása, amennyiben a többi változó állandó, β_i mértékű változást eredményez y-ban.

A modell paramétereit az egyváltozós regresszióhoz hasonlóan a legkisebb négyzetek elvével határozzák meg, ahol a hibanégyzetek összegének minimalizálása a cél.

Becslő és Indikátor változók

Tételezzük fel, hogy „p” apaállat ivadékainak adatai ismertek.
A lineáris modell a következő:



Becslő és Indikátor változók

Tételezzük fel, hogy „p” apaállat ivadékainak adatai ismertek.
A lineáris modell a következő:

$$y_{ij} = \mu + s_i + \varepsilon_{ij}$$

A fenti modellt átírhatjuk lineáris modellé indikátorváltozók használatával.
Az indikátorváltozó:

$$x_{ik} = \begin{cases} 1, \text{sire} = i \\ 0, \text{sire} \neq i \end{cases}$$

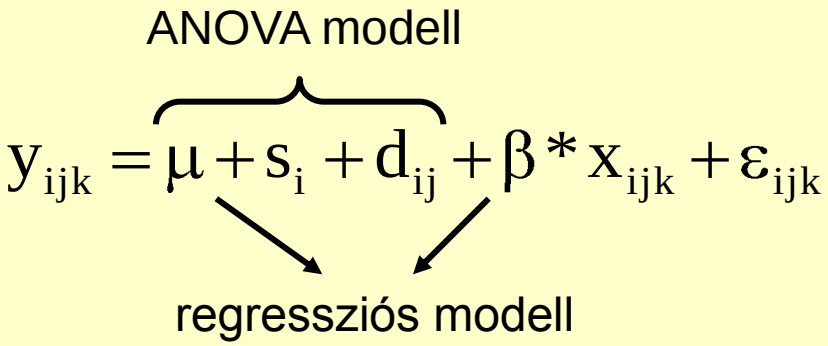
A lineáris modell a következő:

$$y_{ij} = \mu + \sum_{k=1}^{p} \beta_x * x_{ik} + \varepsilon_{ij}$$

Az indikátorváltozókat tartalmazó modelleket **ANOVA**, vagy **varianciaanalízis modelleknek** hívjuk.

Azokat a modelleket, amelyekben a főátlagon kívül nincsenek indikátorváltozók, együtthatóik folytonos, vagy diszkrét értékek lehetnek, **regressziós modelleknek** nevezzük.

Mindkettő az **Általánosított Lineáris Modell (General Linear Model /GLM/)** speciális esete.



Példa: féltestvér/teljestestvér modell, a tulajdonságot a β az életkorra korrigálja.

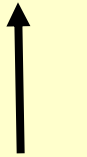
Lineáris modellek mátrixalakja

Legyen adott három változó, és négy mérési vektor.

$$y_i = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

Mátrix alakban: $y = X\beta + \varepsilon$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \quad X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$



A becsülni kívánt paraméterek

Az együtthatók mátrixa

Az apamodell mátrixalakja

A modell: $y_{ij} = \mu + s_i + \varepsilon_{ij}$

Tételezzünk fel három apaállatot. Az elsőnek kettő, a másodiknak egy, a harmadiknak három ivadéka van.

A GLM az alábbiak szerint néz ki:

$$y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{31} \\ y_{32} \\ y_{33} \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \beta = \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \end{bmatrix}$$

Legkisebb négyzetek módszere (Ordinary Least Squares /OLS/)

Ha a hibavektor elemei (a rezidiumok) normál eloszlásúak, az OLS megoldása visszavezethető a Maximum-Likelihood módszerre.

Ha a rezidiumok homogének, és korrelálatlanok, $\sigma^2(e_i)= \sigma_e^2$, $\sigma(e_i,e_j)=0$. Ezért minden rezidiumnak azonos súlya van.

Amennyiben β-t b-vel becsüljük, és az y értékek becslésére ebből $\hat{y}$ értékeket számítunk, a hibavektor a következő lesz:

$$\hat{e} = y - \hat{y} = y - Xb$$

A hibák négyzetösszege a következők szerint számítható:

$$\sum_{i=1}^{n} \hat{e}^2 = \hat{e}^T * \hat{e} = (y - Xb)^T * (y - Xb)$$

y becslése

Ha vesszük a mátrixok deriváltjait, látható, hogy a becslések kielégítik a következő összefüggést:

$$b = \left(X^T * X\right)^{-1} * X^T * y$$

Ez az OLS becslése a β vektornak.

Mivel a hibavektor elemei korrelálatlanok és homogének, a b mátrix elemei közötti kovariancia:

$$V_b = \left(X^T * X\right)^{-1} * \sigma_e^2$$

A σ_e^2 /Var(e)/ a $\mathrm{Var}(e) = \dfrac{\hat{e}^T * \hat{e}}{n - \mathrm{rang}(X)}$ képlettel becsülhető,

ahol a rang(X) az X együttható-mátrix egymástól független oszlopainak számát jelöli.

Példa: Regresszió az origón keresztül

$$y = \beta * X_i + e_i$$

Ahol $\beta = \beta$ és $X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$,
amiből:

$$X^T X = \sum_{i=1}^{n} x_i^2 \qquad X^T y = \sum_{i=1}^{n} x_i * y_i$$

β OLS becslése, valamint a becsült varianciák az alábbiak:

$$b = (X^T * X)^{-1} * X^T * y = \frac{\sum x_i * y_i}{\sum x_i^2} \qquad \sigma^2(b) = (X^T * X)^{-1}$$

Példa: Polinomiális regresszió és interakció

Az általánosított lineáris modell (GLM) egyszerűen kezeli a becslő változók különböző függvényeit, így a becsülni kívánt paraméterek lineárisak maradnak. Pl.: $y=\alpha+\beta_1*f(x)+\beta_2*g(x)+\ldots+e$

Kvadratikus regresszió:

$$y_i = \alpha + \beta_1 * x_i + \beta_2 * x_i^2 + e_i$$

$$\beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

Az interakciókat (pl. ivar x életkor) is hasonlóan kezeli:

$$y_i = \alpha + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \beta_3 * x_{i2} * x_{i2} + e_i$$

Ha az x_1-et állandónak vesszük, x_2 egységnyi változása y-t $\beta_2+\beta_3*x_1$ mértékben változtatja meg.

Hasonlóképpen, az x_1 egységnyi változása y-t $\beta_1+\beta_2*x_2$ mértékben változtatja meg.

$$\beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{11} * x_{21} \\ 1 & x_{21} & x_{22} & x_{21} * x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1} * x_{n2} \end{bmatrix}$$

Hatások figyelembe vétele fix, vagy random hatásként

A lineáris modellek két nagy célnak próbálnak megfelelni:

- a modell paramétereinek becslése,
- a megfelelő varianciák becslése

Például a legegyszerűbb regressziós modellben ($y=\alpha+\beta*x+e$) α és β értékeit, valamint e varianciáját becsüljük. Természetesen $e_i=y_i-(\alpha+\beta*x_i)$ értékeket is tudjuk becsülni.

Fontos, hogy az α ésβ értékek, amelyeket becsülni kívánunk, **állandó hatások** (**fix faktorok**), míg az e_i értékek egy eloszlásból származnak. Az e_i-t ezért **véletlen hatásnak** (**random effektnek**) tekintjük.

Az állandó és véletlen hatások közötti ilyen megkülönböztetés különösen fontos a modellek elemzésekor.
Ha a becsülni kívánt paraméter állandó konstans, akkor az állandó hatás.
Ha a becsülni kívánt paraméter valamilyen eloszlásból származik, és erre az eloszlásra mi következtetéseket kívánunk levonni, akkor véletlen hatás.

Általában **az állandó hatások becsléséről**, és
a véletlen hatások előrejelzéséről beszélünk.

A **„vegyes”** (**„mixed”**) **modellek állandó** és **véletlen hatásokat** egyaránt tartalmaznak.

Példa: apa modell

állandó, vagy véletlen hatás?

$$y_{ij} = \mu + s_i + e_{ij}$$

állandó hatás — véletlen hatás

Attól függ. Amennyiben van 10 apaállatunk, és mi CSAK ennek a 10 apaállatnak az értékére vagyunk kíváncsiak, nem vizsgáljuk a populáció egyéb részét, akkor tekinthetjük **ÁLLANDÓ hatásnak**. Ebben az esetben a μ, s_1-től s_{10}-ig az apaállatok, és σ^2_e becslése szükséges, és a modellt az $y_{ij}=\mu+s_i+e_{ij}$, $\sigma^2(e_{ij}) = \sigma^2_e$ alakban írhatjuk fel.

Példa: apa modell

$$y_{ij} = \mu + s_i + e_{ij}$$

Hasonlóképpen, amennyiben nemcsak ez a 10 apaállat érdekel minket, hanem a kiindulási populációra is szeretnénk következtetéseket levonni, akkor az s_i-t **VÉLETLEN hatásnak** tekintjük. Ebben az esetben a μ-t, és a σ^2_s, σ^2_w varianciákat becsüljük. Mivel az si értékek becsülik (vagy előrejelzik) az i. apaállat tenyészértékét, ezeket szintén becsülni (előrejelezni) szeretnénk. A véletlen hatásként történő figyelembevétel esetén a modellt az $y_{ij}=\mu+s_i+e_{ij}$, $\sigma^2(e_{ij}) = \sigma^2_e$, $\sigma^2(s_i) = \sigma^2_s$ alakban írhatjuk fel.

Általánosított Lineáris Modell (GLS)

Tételezzük fel, hogy a hibaértékek (rezidiumok) eloszlásának nem ugyanolyan a varianciája (pl. heterogenitást mutat). Természetesen a *súlyozott* hibanégyzetek összegét nem szeretnénk minimalizálni, mivel ezek az összegek alacsonyabb varianciánál nagyobb súlyokat kapnának.

Hasonlóképpen, amennyiben a hibaértékek korrelálnak, ezt ugyancsak figyelembe kívánjuk venni (pl. egy megfelelő átalakítással megszüntetni a korrelációt) a négyzetösszegek minimalizálása előtt.

A fenti esetek mindegyikére az OLS megoldás helyett a GLS megoldás szükséges.

A GLS modellben a hibaértékek „e” vektorának kovariancia mátrixát „R” jelöli, ahol $R_{ij} = \sigma(e_i,e_j)$.

A lineáris modell alakja: y = Xb + e, cov(e) = R

A GLS módszerrel „b” becslése β-ra:

$$b = (X^T * R^{-1} * X)^{-1} * X^T * R^{-1} * y$$

A becslő modell variancia-kovariancia értékei az alábbiak szerint számíthatók:

$$V_b = (X^T * R^{-1} * X)^{-1} * \sigma_e^2$$

Példa

A hibaértékek korrelálatlanok, de heterogén eloszlásúak, $\sigma^2(e_i) = \sigma^2_\varepsilon / w_i$.

Például, az i. minta az átlaga az n_i egyednek, ahol $\sigma^2(e_i) = \sigma_e^2/n_i$. Itt $w_i = n_i$

$$R = \mathrm{Diag}(w_1^{-1}, w_2^{-1}, \cdots w_n^{-1})$$

A modell $y_i = \alpha + \beta x_i$ alakú, ahol

$$\beta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Előadás összefoglalása

- Mátrixműveletek (összeadás, kivonás, szorzás, invertálás)
- Lineáris modellek általános alakja
- Legkisebb négyzetek módszere és az általánosított legkisebb négyzetek módszere
- Tényezők figyelembevétele állandó, vagy véletlen hatásként

Előadás ellenőrző kérdései

- Számítsa ki a $\begin{bmatrix} 2 & 4 \\ 5 & 3 \end{bmatrix}$ mátrix inverzét!

- Mi a feltétele két mátrix szorzásának?
- Hogyan néz ki az általánosított lineáris modell (GLM)?
- Mi alapján dönti el, hogy egy tényezőt állandó, vagy változó hatásként vesz figyelembe a modellben?
- Mi a különbség az OLS és a GLS módszerek között?

KÖSZÖNÖM FIGYELMÜKET

Következő
ELŐADÁS/GYAKORLAT CÍME

Tenyészértékbecslés, BLUP

- **Előadás anyagát készítették:**