A biostatisztika alapjai

FOK22-23 Biofizika-2

regresszióanalízis

Liliom Károly

biostatisztika előadások tematikája

1. leíró statisztika (adatok jellemzése)
2. hipotézisvizsgálatok (adatok összehasonlítása)
3. korreláció és regresszióanalízis

Javasolt tankönyv:

- **Orvosi Biofizikai Gyakorlatok, Statisztika pótfejezet** – Medicina Kiadó, 2017

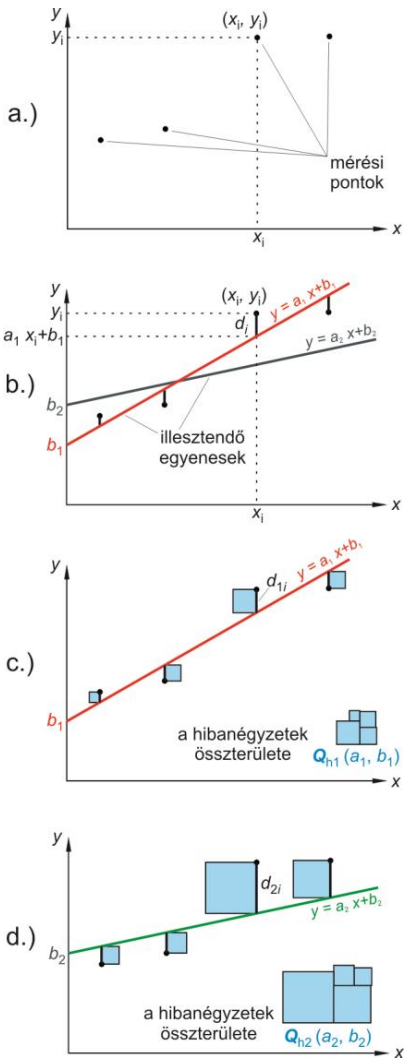
Javasolt olvasmányok:

- Herényi Levente: Statisztika és Informatika, Medicina Kiadó 2016

- Harvey Motulsky: Intuitive Biostatistics – A Nonmathematical Guide to Statistical Thinking, Oxford University Press
- Nature Collection: Statistics for Biologists

Ilyenkor az x illetve az y értékeket logaritmálás nélkül, a tengelyek logaritmikus értékbeosztásai szerint ábrázoljuk (lásd 12. alsó ábrák).

(Megjegyzés. A számítógépek világában az adatok ábrázolása, illetve azok transzformációja egyszerűen elvégezhető. Egyetlen parancs kiadásával a tengelyek átskálázhatók vagy szükség esetén logaritmikus beosztásúvá alakíthatók.)



13. ábra. *A mérési pontokra legjobban illeszkedő egyenes keresése.*

lineáris regresszió
linear regression
lineare Regression

LINEÁRIS REGRESSZIÓ

A legegyszerűbb görbéhez, az egyeneshez papíron, ceruzával, vonalzóval szubjektív módon könnyen eljuthatunk, de a mérési pontok hibájából származó bizonytalanságainkat, kételyeinket a görbekihúzás korábbi kvalitatív elveivel nem tudjuk mindig eloszlatni. Ezért tovább él bennünk az a kérdés, hogy **hogyan találjuk meg a pontjainkra** valóban **legjobban illeszkedő egyenest**.

Tudjuk, hogy az egyenes egyenlete

$$y = a \cdot x + b\,, \tag{15}$$

amelyben a az egyenes meredeksége, b pedig a tengelymetszet, azaz y-nak azon értéke, amelynél az egyenes az y tengelyt az $x = 0$ helyen metszi. E **két paraméter az egyenest egyértelműen jellemzi**. A feladat tehát az, hogy határozzuk meg a mérési pontjainkra legjobban illeszkedő egyenes a^* és b^*-gal jelölt paramétereit.
Első lépésként azt kell megmondanunk, hogy **mit jelent** az, hogy egy egyenes a **legjobban illeszkedő**.

Tegyük fel, hogy az x és y mennyiségekre vonatkozóan rendelkezünk 4 mérési ponttal (lásd 13. a. ábra). Feltesszük továbbá azt is, hogy a mérés során az x_i-k „pontos” (hiba nélküli), előre beállított értékek, tehát (mérési) hiba csak az y_i-ket terheli.

Húzzunk a mérési pontjainkon át egy tetszőleges — a_1, b_1 paraméterekkel adott — egyenest ($y = a_1 \cdot x + b_1$) és határozzuk meg a pontoknak az egyenestől függőlegesen mért távolságait (lásd 13. b ábra, függőleges vonalak).

Egy kiszemelt (x_i, y_i) pont x tengelytől mért távolságát az y_i koordináta adja meg, ugyanakkor az x_i koordinátához tartozó egyenesbeli pont x tengelytől mért távolságát az egyenes egyenletébe való behelyettesítéssel az $(a_1 \cdot x_i + b_1)$ érték adja meg. A kettő különbsége **a pont és egyenes függőlegesen mért távolsága** $d_{1i} = (y_i - (a_1 \cdot x_i + b_1))$. Ha a pont az egyenes fölött van, akkor ez a kifejezés pozitív, ha alatta, akkor negatív. Az így definiált távolságot a többi pontra hasonlóképpen kiszámíthatjuk, majd vegyük ezek négyzeteit.

A tapasztalati szórás bevezetésénél szereplő négyzetes kifejezés (6) mintájára adjuk össze az összes pontra kiszámított ilyen **távolságnégyzetet** (lásd 13. c. ábra, kék területek) és az összeget jelöljük Q_{h1}-gyel. Ha most behúzunk egy másik — mondjuk a_2, b_2 paraméterekkel megadott — egyenest, akkor ugyancsak kiszámíthatók ettől a másik egyenestől mért függőleges távolságok is (d_{2i}), és végeredményként egy másik Q_{h2}-vel jelölt négyzetes összeget kapunk (lásd 13. d. ábra). Megfigyelhetjük, hogy a pontok (egyenestől mért) „szétszórtsága” akkor nagyobb, amikor Q_h is nagyobb ($Q_{h2} > Q_{h1}$).

Mivel az (x_i, y_i) pontok változatlanok, ezért a Q_h változását csak az a, b paraméterek változása okozza. A fenti eljárással (hozzárendeléssel) egy olyan függvényt adtunk meg, ahol a két független változó a és b, a függő változó pedig a Q_h:

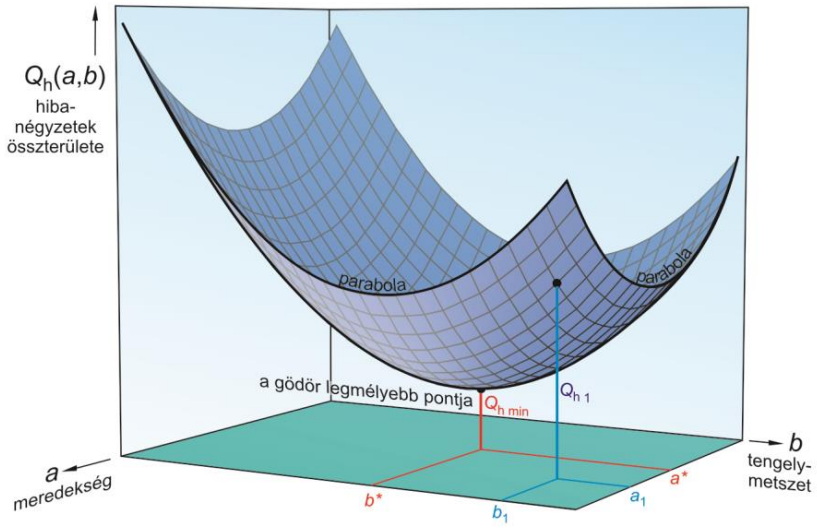
$$Q_h(a, b) = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2\,. \tag{16}$$

A kapott $Q_h(a, b)$ kétváltozós függvény — a távolságnégyzetek miatt — mindkét változójától négyzetesen függ, ami azt jelenti, hogy egy olyan gödörhöz hasonló felülettel reprezentálható, amelynek mindkét tengelyirányú síkmetszete (a = állandó, illetve b = állandó) parabola (14. ábra).

Néhány további feltétel teljesülése esetén (például a mérési hibák az egyes pontokban legyenek egymástól függetlenek) a pontokhoz **legjobban illeszkedő** egyenes az, amelyre nézve az említett **távolság négyzetösszeg**, azaz Q_h **minimális**.

Tehát azt az (a^*, b^*) értékpárt keressük, ahol az imént definiált függvény (16) értéke a legkisebb. Még egyszerűbben: meg kell keresnünk a gödör (14. ábra) legmélyebb pontjának (a^*, b^*) koordinátáit. A fenti módon illesztett egyenest **regressziós egyenesnek**, az illesztési módszert **legkisebb négyzetek módszerének** vagy általánosabban **lineáris regressziónak** szokás nevezni.

Megjegyezzük, hogy a **regresszió** szó a „visszahatásra”, utal, azaz a mérési pontokból a mennyiségek között fennálló kapcsolatra való következtetést fejezi ki. (A „kapcsolat” általánosan továbbra sem jelent feltétlenül oksági kapcsolatot.)



14. ábra. *A hiba függése az* a *és* b *paraméterektől. A gödör legmélyebb pontján a hibanégyzetek összege minimális.*

A minimum keresés elvégzése után azt kapjuk, hogy

$$a^* = \frac{Q_{xy}}{Q_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \text{ vagy } a^* = \frac{s_{xy}^2}{s_x^2}, \qquad (17)$$

$$b^* = \bar{y} - a^*\bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - a^* \frac{\sum_{i=1}^{n} x_i}{n}, \qquad (18)$$

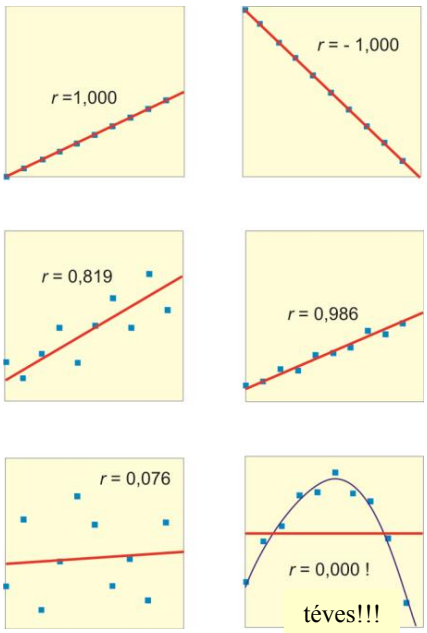
ahol Q_{xx} , Q_{xy} a (6)-ban bevezetett jelölésnek felel meg,

— $s_{xy}^2 = Q_{xy}/(n-1)$ a **kovariancia**,

— s_x^2 az x **varianciája**,

— $\bar{x}$, $\bar{y}$ pedig a megfelelő **átlagokat** jelöli.

A fenti képletek alapján tetszőleges (x_i, y_i) mért érték párok esetén megadható a legjobban illeszkedő egyenes ($y = a^* x + b^*$), még akkor is, ha a pontok szemmel láthatóan nem egyenesen, hanem valamilyen görbe mentén helyezkednek el.



15. ábra. *Néhány példa a korrelációs együttható értékére.*

korrelációs együttható
correlation coefficient
Korrelationskoeffizient

Mivel szubjektív megítélés alapján nagyon kétséges annak eldöntése, hogy a mérési pontok mennyire jól illeszkednek a regressziós egyenesre, ezért célszerű meghatározni a **korrelációs együtthatót**:

$$r = \frac{Q_{xy}}{\sqrt{Q_{xx} \cdot Q_{yy}}} = \frac{s_{xy}^2}{s_x s_y}, \qquad (19)$$

ahol a használt jelölések ugyanazok, mint a (17) kifejezésben.

A korrelációs együttható a változók közötti **kapcsolat szorosságát jellemzi**, r értéke +1 és –1 között változhat. Pozitív értékeihez pozitív, negatív értékeihez negatív meredekségű egyenes tartozik. Ha a mérési pontok jól megközelítik a regressziós egyenest, az $|r|$ értéke közel lesz az 1-hez (pl.: r = 0,9860). Ha az egyenes minden egyes ponton átmegy, akkor $|r| = 1$, egyébként r annál jobban megközelíti a nullát, minél inkább eltérnek a pontok az illesztett egyenestől (lásd a 15. ábrát és egy konkrét példát a 7. megjegyzésben).

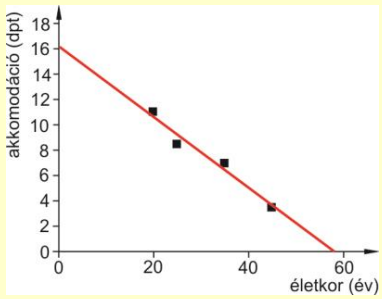
Nem beszéltünk még arról, hogy **milyen esetekben és milyen célra** használhatjuk a lineáris regressziót. Itt egy fontos **szempontot** kell figyelembe vennünk. Nevezetesen azt, hogy a változók (x, y) között **van-e** valamilyen „oksági” összefüggést leíró **modell**, amelynek paraméterei fizikai értelemmel bírnak, vagy csak azért illesztünk egyenest, hogy az a szükséges pontossággal reprezentálja mérési adatainkat. (Természetesen amennyiben a modell nem ismeretes, az még nem jelenti azt, hogy az oksági viszony nem állhat fenn.)

Az első esetre többek között példa lehet a folyadékok törésmutatójának koncentrációtól való függése ($n = n_0 + k\,c$) (lásd 4. REFRAKTOMETRIA című fejezet), és a folyadékok optikai denzitásának koncentrációtól való függése ($\lg (J_o / J) = \varepsilon\, c\, x$) (lásd 6. FÉNYABSZORPCIÓ című fejezet). Ilyenkor a lineáris regresszió paraméterei **extrapolációra** (a mérési tartományon kívüli becslésre) is használhatók. Az összefüggések (törvények) érvényességi tartományait azonban itt is figyelembe kell vennünk.

A második esetben a paramétereket csak **interpolációra** (a mérési tartományon belüli becslésre) használhatjuk és még 1-hez közel eső korrelációs együttható sem ad alapot arra, hogy **oksági kapcsolatról beszéljünk** (lásd 7. megjegyzés).

7. megjegyzés:
Az alábbi táblázatban a szem akkomodációs képességét tüntettük fel az életkor függvényében

életkor (év)	20	25	35	45
akkomodáció (dpt)	11	8,5	7	3,5



16. ábra. *A szem akkomodációja az életkor függvényében. Regressziós egyenes illesztése a mérési pontokra.*

Az ábrázolás és a lineáris regresszió elvégzése után az illesztett egyenes paraméterei:

$a^* = -0{,}28; \quad b^* = 16{,}2.$

A korrelációs együttható:

$r = -0{,}98.$

Mivel nem ismerünk olyan matematikai modellt, törvényszerűséget, ami a két mennyiséget összekapcsolja, ezért a paramétereket csak interpolációs becslésre használhatjuk fel. Például megkaphatjuk az akkomodáció becsült értékét 40 éves korban:

$-0{,}28 \cdot 40 + 16{,}2 \approx 5$ (dpt)

ami a táblázatból egyébként hiányzik. (Tudjuk azonban, hogy még egészen fiatal korban sincs 16 dpt körüli akkomodáció.)