

Regresszió és korreláció

regression:
visszatérés, hátrálás;
visszafordulás

correlation:
viszony, összefüggés,
kölsönösség

KAD 2014.11.10

1

Regresszió és korreláció

(visszatérés, hátrálás; visszafordulás) (viszony, összefüggés, kölsönösség)

Gyakorlati megközelítés (pl.1)

mennyi fehérje van a vérplazmában?

(db, mol, g, ...)

mekkora a vérplazma fehérjekoncentrációja?

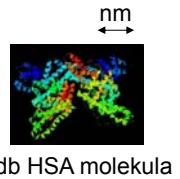
(db/L, mol/L, g/L)

Nephrosis (súlyos vesebetegség) esetén értéke erősen lecsökken

direkt módszer: megszámolni egy adott térfogatban levő fehérje molekulák számát(?)

közvetett módszer:

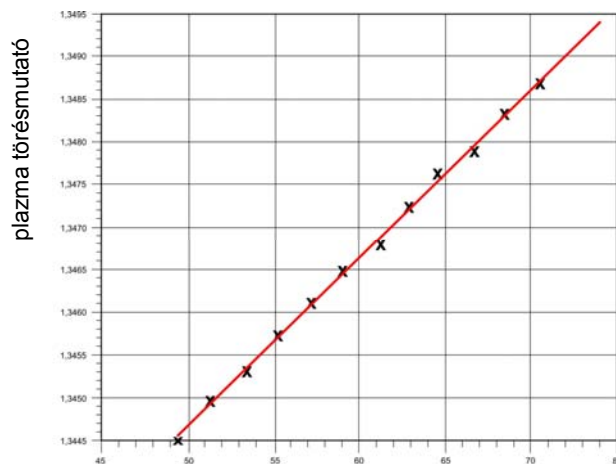
keresni egy olyan (könnyen) mérhető fizikai mennyiséget, amely szigorúan monoton kapcsolatban van a megismerni kívánt mennyiséggel (legegyszerűbb ilyen függvény ...)



2

észrevétel:

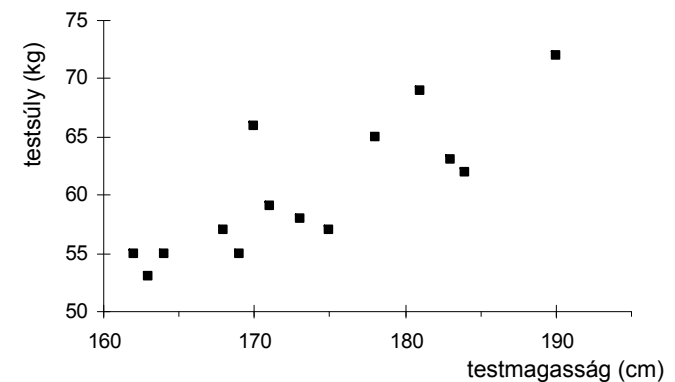
a vérplazmában a fény lassabban halad, ha sok benne a fehérje (magas a fehérjekoncentráció), azaz nagyobb a törésmutatója (determinisztikus kapcsolat, de: mérési hiba mindig van)



plazma
fehérje
koncentráció
(g/L)

3

(pl.2) E2 csoport (1994.09) tagjainak adatai (összetartozó értékpárok)



| cm | kg |
|-----|----|
| 162 | 55 |
| 163 | 53 |
| 164 | 55 |
| 168 | 57 |
| 169 | 55 |
| 170 | 66 |
| 171 | 59 |
| 173 | 58 |
| 175 | 57 |
| 178 | 65 |
| 181 | 69 |
| 183 | 63 |
| 184 | 62 |
| 190 | 72 |

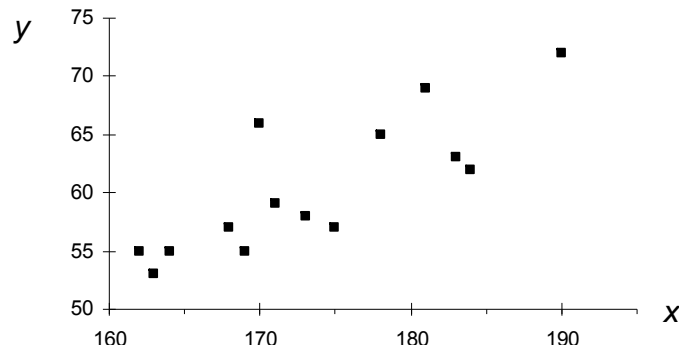
milyen tendenciát látunk?

ezen előadásfél hosszú címe:
Ugyanabban a csoportban felvett többféle kvantitatív változó közötti kapcsolat elemzése. Korreláció, lineáris regresszió, a korrelációs koefficiens fogalma

4

A korrelációs számítás két véletlen számszerű változó szimmetrikus kapcsolatával foglalkozik

akkor beszélünk korrelációs kapcsolatáról az x és y véletlen változók között, ha vagy kis x értékekhez kis y értékek, nagy x értékekhez nagy y értékek (**pozitív** kapcsolat), vagy pedig kis x értékekhez nagy y értékek és nagy x értékekhez kis y értékek (**negatív** kapcsolat) tartoznak

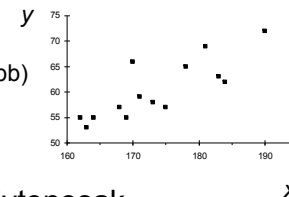


itt: pozitív korreláció

5

Regressziós megközelítés

függvénykapcsolatot keresünk egy (vagy több) **független változó** (x) és egy **függő változó** (y) között



feltételezések: x és y számszerűek és folytonosak,
 y valószínűségi változó (értékét nem csak a magyarázó változók, hanem a véletlen is befolyásolja)

A regressziós modell rögzíti a **függvény típusát**:

lineáris $y = (ax + b) + h$ (a : meredekség, b : tengelymetszet)

polinomiális $y = a + b_1x + b_2x^2 + \dots + b_nx^n + h$

exponenciális $y = ab^x h$

hatványfüggvényes $y = ax^b h$

és azt, hogy **hogyan hat a véletlen** a függő változóra:

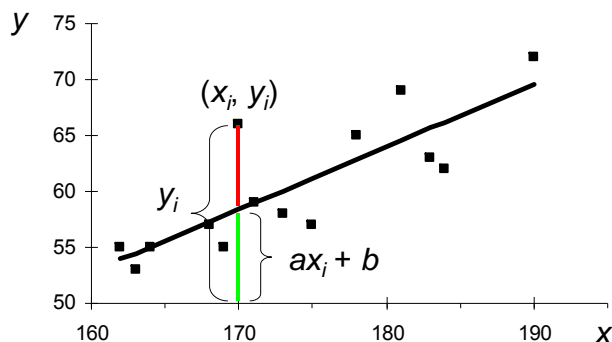
Lineáris és polinomiális esetben a független változók értékétől független, **additív** ($+h$) hibával, exponenciális és hatvány esetben **multiplikatív** ($\cdot h$) hibával.

6

A legegyszerűbb regressziós modell a lineáris regresszió

lineáris függvény: $y = (ax + b) + h$

$h_i = y_i - (ax_i + b)$ Ha a pont (x_i, y_i) az egyenes fölött van.
(Hogyan írható fel, ha alatta van?)



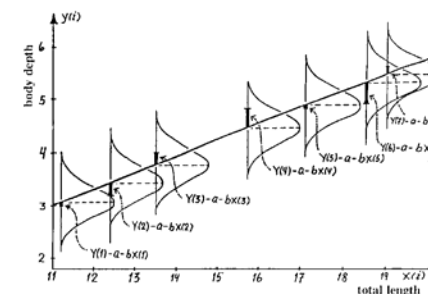
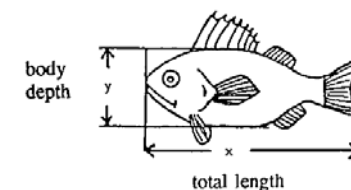
| | x_i | y_i |
|----|-------|-------|
| 1 | 162 | 55 |
| 2 | 163 | 53 |
| 3 | 164 | 55 |
| 4 | 168 | 57 |
| 5 | 169 | 55 |
| 6 | 170 | 66 |
| 7 | 171 | 59 |
| 8 | 173 | 58 |
| 9 | 175 | 57 |
| 10 | 178 | 65 |
| 11 | 181 | 69 |
| 12 | 183 | 63 |
| 13 | 184 | 62 |
| 14 | 190 | 72 |

Legjobb egyenes: hibák négyzetösszege a lehető legkisebb (**legkisebb négyzetek** módszere)

7

Az alkalmazhatóság feltételei

1. x és y között lineáris a kapcsolat.
2. A mintán belüli megfigyelési pontok egymástól függetlenek.
3. Minden rögzített x értékre az y értékek eloszlása normális.
4. Az y értékek eloszlása minden x értékre ugyanazzal a varianciával rendelkezik.
5. Az x értékeket „hiba nélkül” lehet mérni.



8

a (négyzetes) hibafüggvény:

$$Q_h(\dots) = \sum_{i=1}^n [y_i - (ax_i + b)]^2 \quad \text{mi(k) a független változó(k)?}$$

a és b

milyen a függvénykapcsolat *a*-ra és *b*-re nézve?

mindegyik változóban négyzetes a kapcsolat

milyen függvénnyel ábrázolhatók?

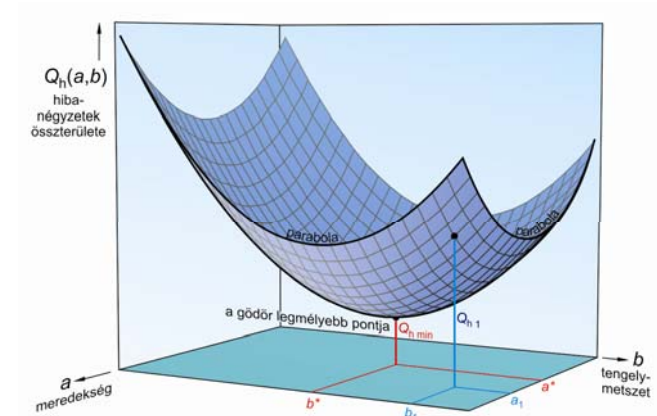
különböző tágasságú parabolákkal

minimummal vagy maximummal rendelkeznek?

grafikonjuk minimummal rendelkező parabola

9

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

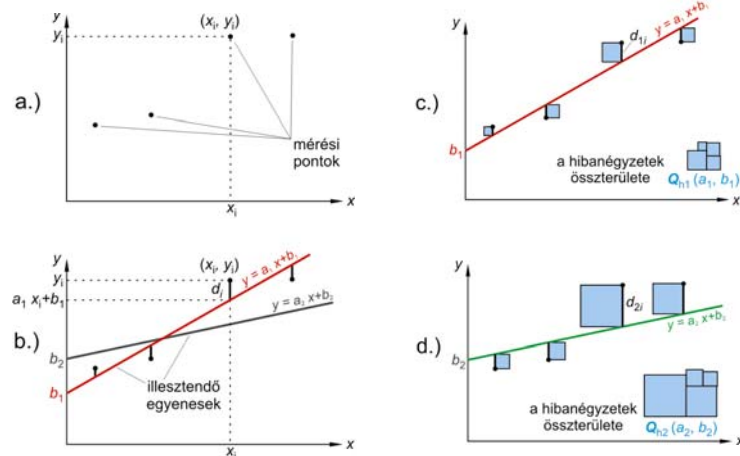


Gyak. jegyzet 2a fej. 14. ábra

10

A mérési pontokra legjobban illeszkedő egyenes ($y = ax + b$) keresése

*a: meredekség
b: tengelymetszet*



11

Gyak. jegyzet 2a fej. 13. ábra

Lineáris regresszió

$$Q_h(a, b) = \sum_{i=1}^n [y_i - (ax_i + b)]^2 \quad \text{hibafüggvény minimalizása}$$

megoldási lehetőségek:

1. teljes négyzetté kiegészítés

$$\text{pl. } y = x^2 - 6x + 14 = (x-3)^2 + 5, \text{ minimum } x = 3\text{-nál}$$

2. differenciálszámítás

differenciálhányados: az érintő iránytangense

szélsőérték keresés: ahol a görbének minimuma (vagy maximuma) van, ott az érintő iránytangense zérus

a szerinti és *b* szerinti differenciálhányadosok zérusok, 2 egyenlet, 2 ismeretlen (2 ismeretlenes lineáris egyenletrendszer)

12

a „legjobb” meredekség:

$$(y = ax + b)$$

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

a „legjobb” tengelymetszet:

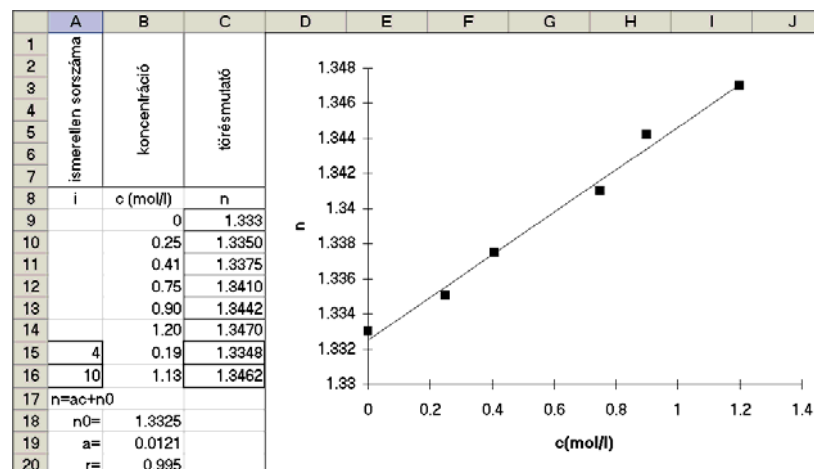
$$b^* = \bar{y} - a^* \cdot \bar{x} = \frac{\sum_{i=1}^n y_i}{n} - a^* \frac{\sum_{i=1}^n x_i}{n}$$

ahol $s_{xy}^2 = \frac{Q_{xy}}{n-1}$: kovariancia

13

Példa: refraktometria

(ismeretlen koncentráció meghatározása kalibrációs egyenes segítségével)



14

Milyen a pontok illeszkedése a regressziós egyeneshez?

ehhez a **korrelációs számítás** nyújt segítséget
(két véletlen változó szimmetrikus kapcsolatával foglalkozik)

a változók közötti kapcsolat erősségét vizsgálja (van erős és gyenge korreláció)

korrelációs együttható
(Pearson-féle)

$$r = \frac{Q_{xy}}{\sqrt{Q_{xx} \cdot Q_{yy}}} = \frac{s_{xy}^2}{s_x s_y}$$

a számláló megegyezik a regressziós egyenes meredekségének számlálójával (a nevező mindkét esetben pozitív)

$a^* = \frac{Q_{xy}}{Q_{xx}}$ \rightarrow pozitív meredekség: $r > 0$ (pozitív korreláció)
negatív meredekség: $r < 0$ (negatív korreláció)

$$-1 \leq r \leq 1$$

15

$$-1 \leq r \leq 1$$

$$0 \leq r^2 \leq 1$$

korrelációs együttható
(Pearson-féle)

meghatározottsági együttható
(coefficient of determination)

legsorosabb, „függvényszerű” kapcsolat: 1
a kapcsolat hiánya (korrelálatlanság): 0

Ez vagy függetlenséget jelent, vagy pedig azt, hogy a választott mérőszám nem alkalmas a szóban forgó kapcsolat mérésére.

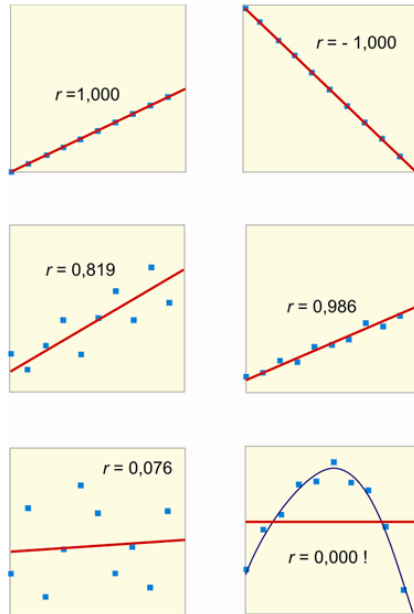
rangsámokból számolt korrelációs együttható:
Spearman-féle rangkorrelációs együttható

szintén csak az adatok sorrendjét (de számértékét nem) használja fel az előbbinél egyszerűbben számolható:

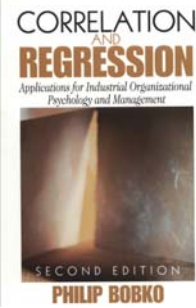
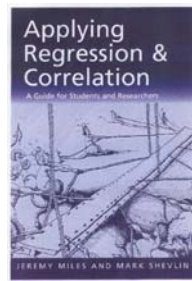
Kendall-féle rangkorrelációs együttható

16

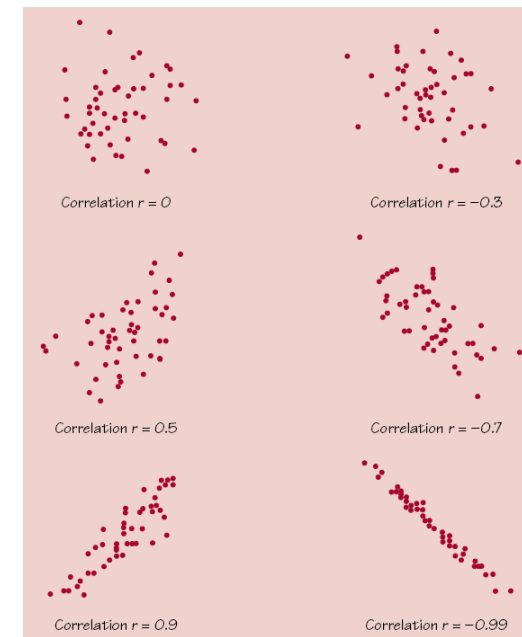
Példák:



Gyak. jegyzet 2a fej. 15. ábra



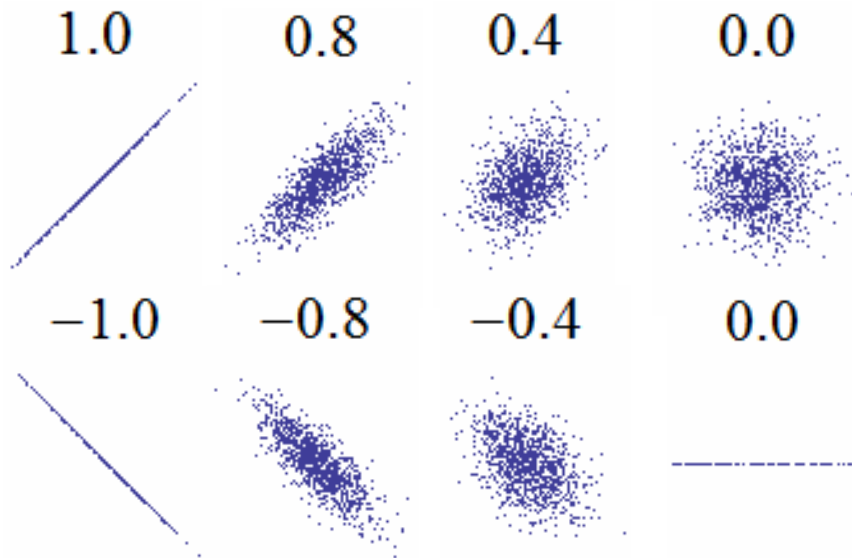
Pont-
diagramok



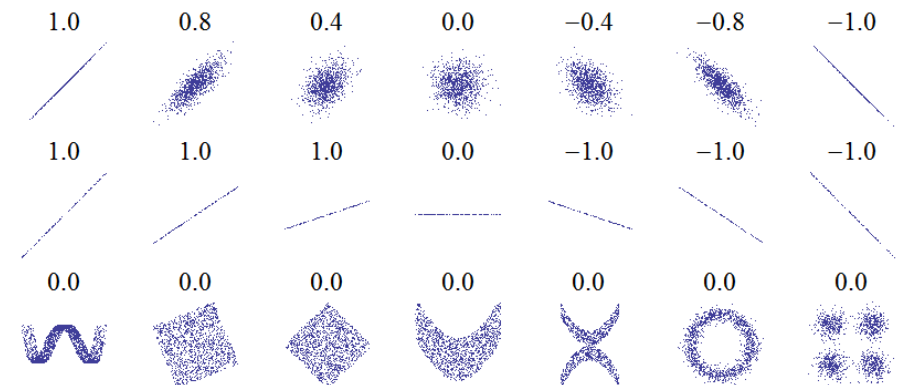
<http://www.cshgreenwich.org/faculty/byrne/chapter%203.pdf>

18

Példák korrelációs együtthatókra



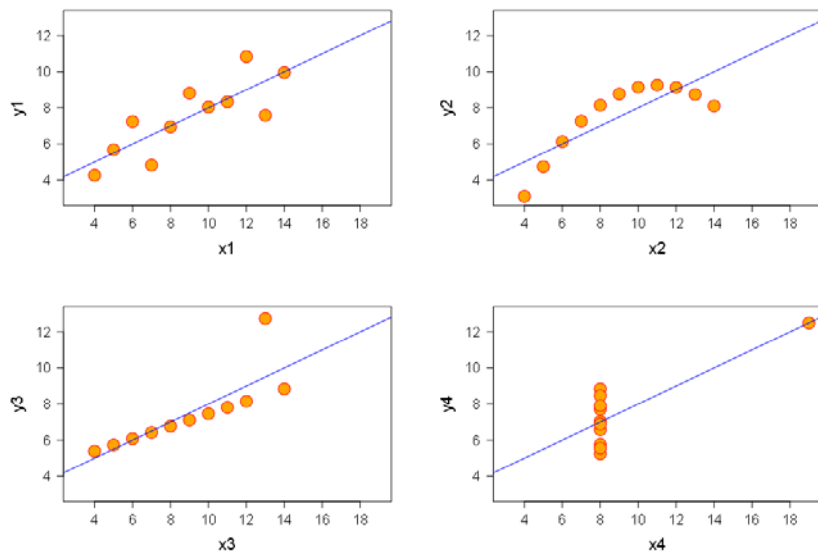
http://en.wikipedia.org/wiki/Image:Correlation_examples.png



http://en.wikipedia.org/wiki/Image:Correlation_examples.png

20

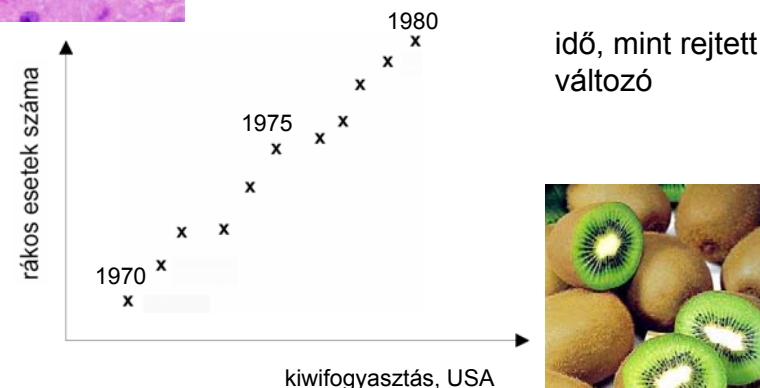
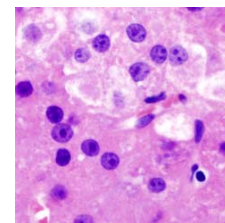
Extrém példa: $r = 0.816$, $y = 0.5x + 3$ (Anscombe)



http://en.wikipedia.org/wiki/Anscombe%27s_quartet

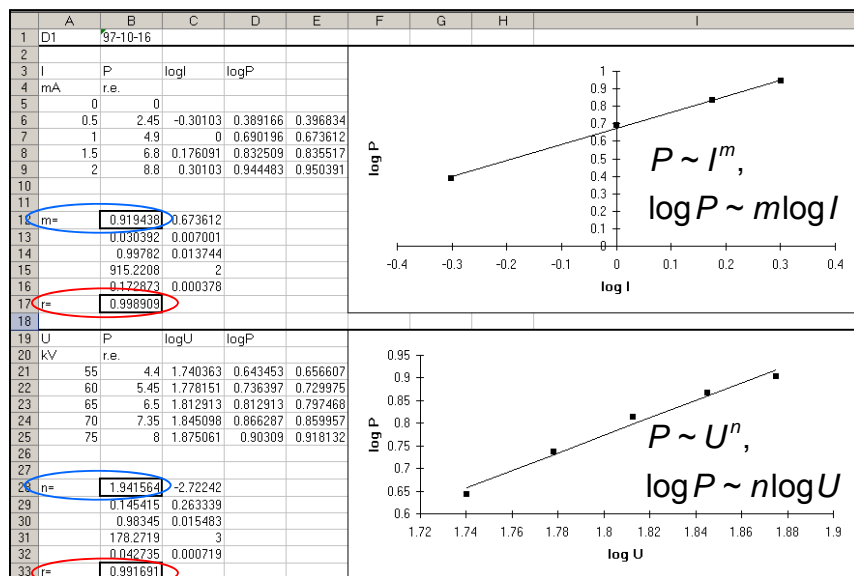
21

A korreláció jelenléte
nem (feltétlenül) jelent
okási kapcsolatot



idő, mint rejtett
változó

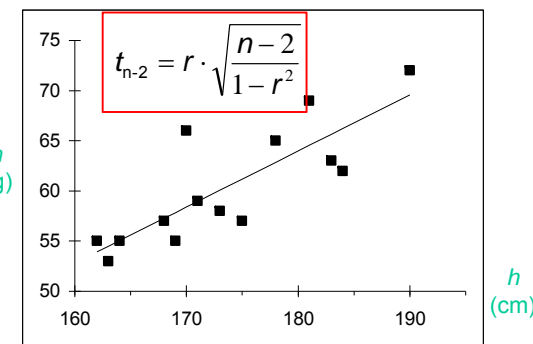
Példa: Hatványfüggvényes regresszió visszavezetése lineáris regresszióra. Röntgenső sugárteljesítményének mérése



Korrelációs t-próba

| cm | kg | E2 |
|----------------|-------|-------|
| 162 | 55 | 53.93 |
| 163 | 53 | 54.49 |
| 164 | 55 | 55.05 |
| 168 | 57 | 57.28 |
| 169 | 55 | 57.84 |
| 170 | 66 | 58.39 |
| 171 | 59 | 58.95 |
| 173 | 58 | 60.07 |
| 175 | 57 | 61.19 |
| 178 | 65 | 62.86 |
| 181 | 69 | 64.54 |
| 183 | 63 | 65.65 |
| 184 | 62 | 66.21 |
| 190 | 72 | 69.56 |
| a= | | |
| 0.558 -36.5096 | | |
| 0.113 19.66358 | | |
| 0.67 3.492297 | | |
| 297.1 146.3537 | | |
| b= | | |
| r= | 0.819 | |
| n= | 14 | |
| t= | 4.935 | |

Van-e kapcsolat a két
mennyiség között?



H_0 : nincs kapcsolat

Határozzuk
meg pontos-
sabban!

$$|t| = 4.935 > t_{12, \text{krit}(0,05)} = 2.18 \Rightarrow H_0 \text{ hamis } (p < 0.05)$$

24