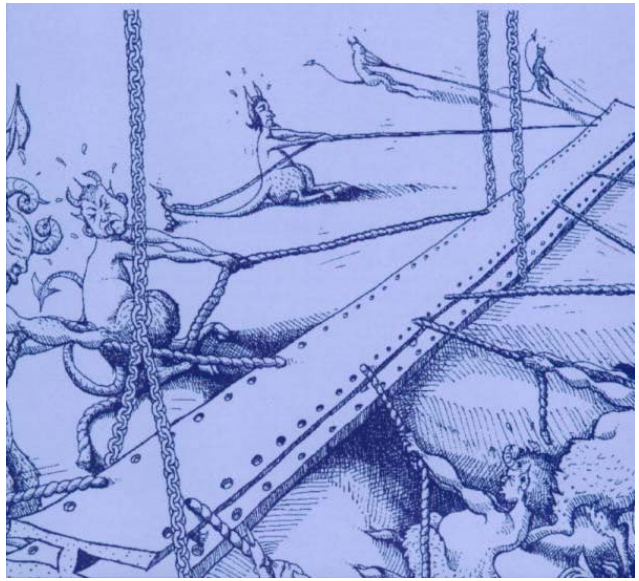


Regresszió és korreláció

regression:
visszatérés, hátrálás;
visszafordulás

correlation:
viszony, összefüggés,
kölcsönösség

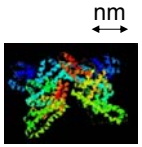
Regresszió és korreláció

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

Gyakorlati megközelítés (pl.1)

mennyi fehérje van a vérplazmában?
(db, mol, g, …)

mekkora a vérplazma fehérjekoncentrációja?
(db/L, mol/L, g/L)



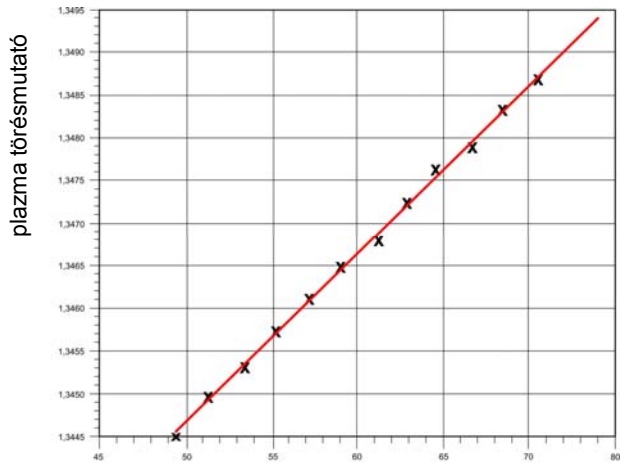
1 db HSA molekula

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 ...)

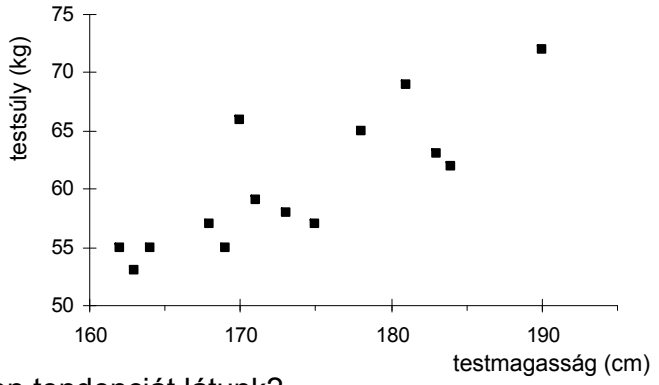
észrevétel:
a vérplazmában a fény lassabban halad, ha sok benne a fehérje (magas a fehérjekoncentráció), azaz nagyobb a törésmutatója (determinisztikus kapcsolat, de: mérési hiba mindig van)



a változások determinisztikus és sztochasztikus része mindig együtt fordul elő

(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

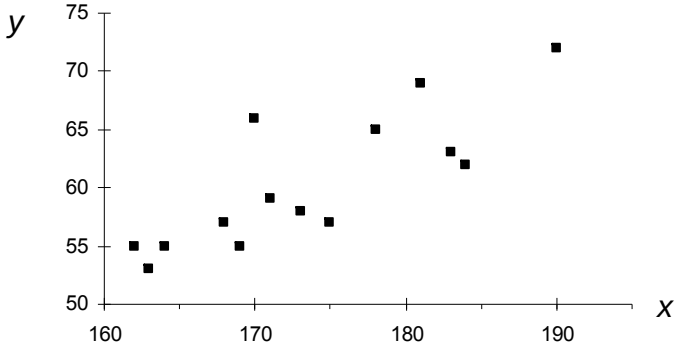


milyen tendenciát látunk?

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

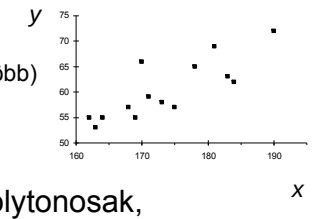
A **korrelációszámítás** két véletlen számszerű változó **szimmetrikus** kapcsolatával foglalkozik

akkor beszélünk korrelációs kapcsolatról 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

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 $y = a + b_1x + b_2x^2 + ... + b_nx^n + h$

exponenciális $y = ab^x\ h$

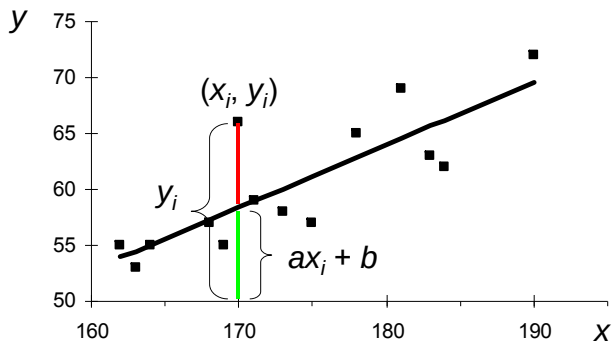
hatványfüggvényes $y = ax^b\ h$

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** (·*h*) hibával.

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?)

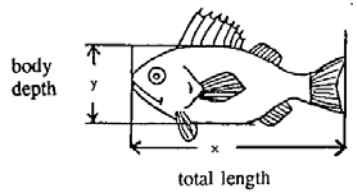


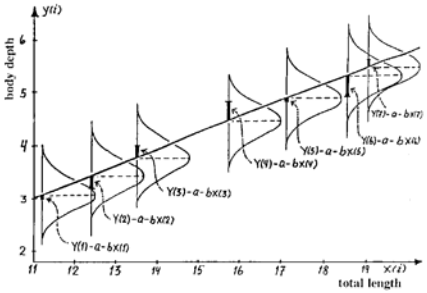
	x_i	y_i
1	162	55
2	163	53
3	164	55
4	168	57
5	169	55
6	170	66
7	171	59
8	173	58
9	175	57
10	178	65
11	181	69
12	183	63
13	184	62
14	190	72

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

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.





http://www.fao.org/docrep/w5449e/w5449e04.htm

a (négyzetes) **hibafüggvény**:

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

a és *b*

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

mindegyik változóban négyzetes a kapcsolat

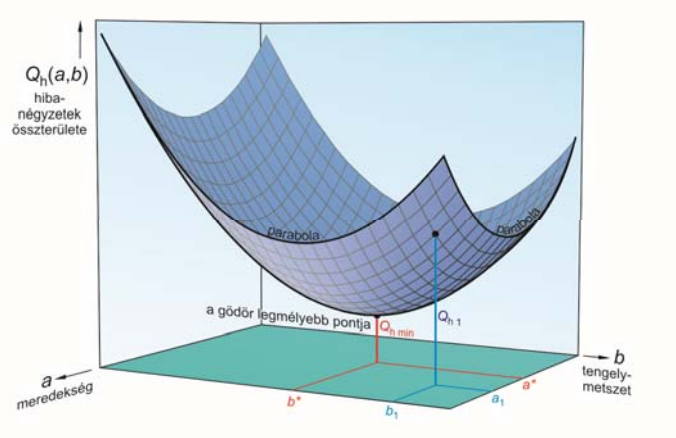
milyen függvénnyel ábrázolhatók?

különböző tágasságú parabolákkal

minimummal vagy maximummal rendelkeznek?

grafikonjuk minimummal rendelkező parabola

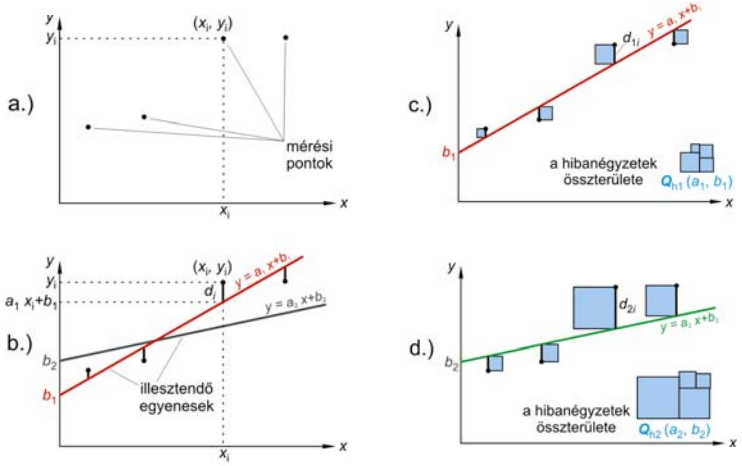
$$Q_h(a, b) = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 = \min.$$



Gyak. jegyzet 2a fej. 14. ábra

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

a: meredekség
b: tengelymetszet



Gyak. jegyzet 2a fej. 13. ábra

Lineáris regresszió

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

megoldási lehetőségek:

1. teljes négyzetté kiegészítés

   pl. $y = x^2-6x+14 = (x-3)^2+5$, minimum x = 3-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)

a „legjobb” meredekség: ($y = ax + b$)

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

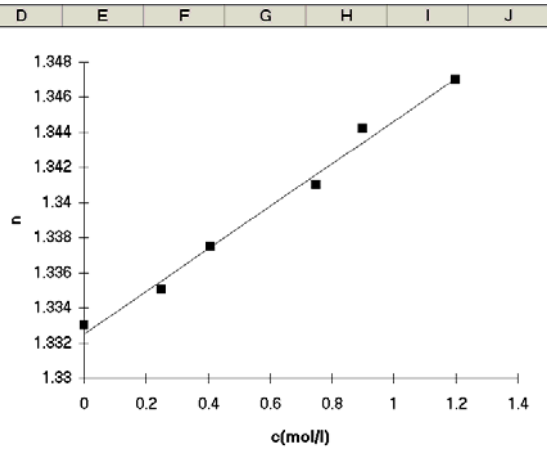
a „legjobb” tengelymetszet:

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

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

Példa: refraktometria
(ismeretlen koncentráció meghatározása
kalibrációs egyenes segítségével)

	A	B	C
1	ismeretlen sorszáma	koncentráció	törésmutató
2			
3			
4			
5			
6			
7			
8	i	c (mol/l)	n
9		0	1.333
10		0.25	1.3350
11		0.41	1.3375
12		0.75	1.3410
13		0.90	1.3442
14		1.20	1.3470
15	4	0.19	1.3348
16	10	1.18	1.3462
17	n=ac+n0		
18	n0=	1.3325	
19	a=	0.0121	
20	r=	0.995	



Milyen a pontok illeszkedése a regressziós egyeneshez?

ehhez a **korrelációszámítás** 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)

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

negatív meredekség: $r < 0$ (negatív korreláció)

$$-1 \le r \le 1$$

$$-1 \le r \le 1 \qquad 0 \le r^2 \le 1$$

korrelációs együttható (Pearson-féle)

meghatározottsági együttható (coefficient of determination)

azt az arányt jelöli, amennyire a regressziós egyenlet meg tudja magyarázni a függő változó varianciáját
pl. $r^2 = 0{,}25$ azt jelenti, hogy a regressziós egyenlet a függő változó varianciájának 25%-át tudja megmagyarázni

legszorosabb, „**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

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\frac{\sum x_i y_i}{n} - \left(\frac{\sum x_i}{n}\right)\left(\frac{\sum y_i}{n}\right)}{\sqrt{\left(\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2\right)\left(\frac{\sum y_i^2}{n} - \left(\frac{\sum y_i}{n}\right)^2\right)}}$$

rangszámokból számolt korrelációs együttható:

Spearman-féle rangkorrelációs együttható (r_s)

(Pearson-féle korreláció a rangszámok között)

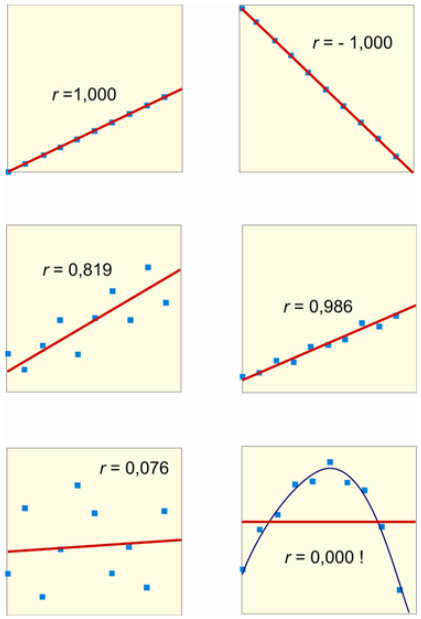
$$r_S = \frac{\sum_{i=1}^{n}(R_{xi}-\bar{R_x})(R_{yi}-\bar{R_y})}{\sqrt{\sum_{i=1}^{n}(R_{xi}-\bar{R_x})^2\sum_{i=1}^{n}(R_{yi}-\bar{R_y})^2}} \qquad \tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

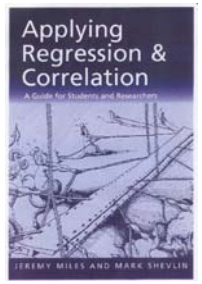
szintén csak az adatok sorrendjét használja fel az előbbinél egyszerűbben számolható:

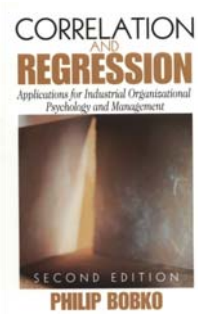
Kendall-féle rangkorrelációs együttható (τ)

(a pozitív és a negatív kapcsolatok arányának a különbségét számolja ki)

Példák:

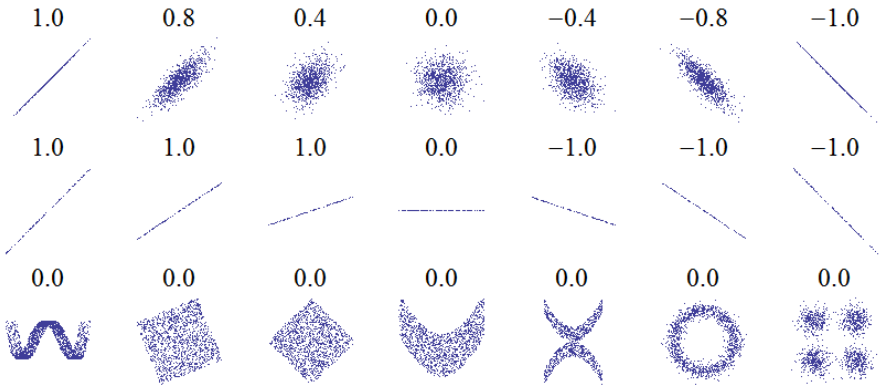






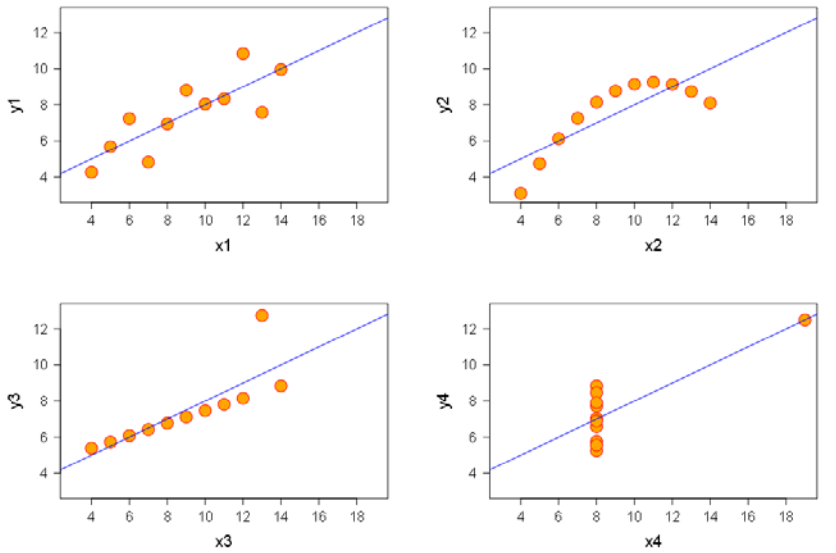
Gyak. jegyzet 2a fej. 15. ábra

Példák korrelációs együtthatókra

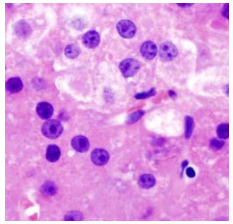


http://en.wikipedia.org/wiki/Image:Correlation_examples.png

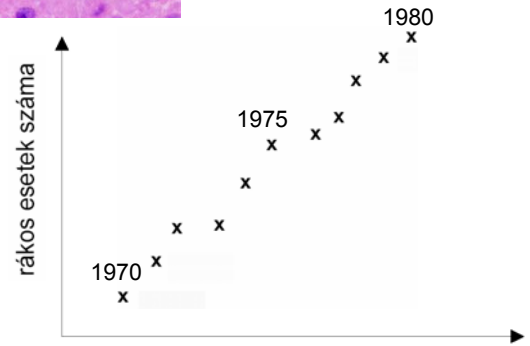
Extrém példa: $r = 0.816$, $y = 0.5x + 3$ (Anscombe)



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



A korreláció jelenléte **nem** (feltétlenül) jelent **oksági kapcsolatot**

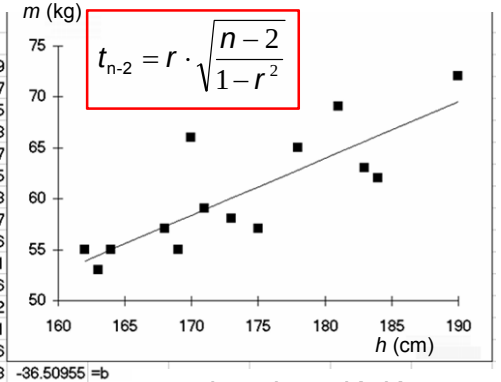


idő, mint rejtett változó

Korrelációs t-próba

Van-e kapcsolat a két mennyiség között?

magasság (cm)	tömeg (kg)		
162	55	53.929	
163	53	54.487	
164	55	55.045	
168	57	57.278	
169	55	57.837	
170	66	58.395	
171	59	58.953	
173	58	60.07	
175	57	61.186	
178	65	62.861	
181	69	64.536	
183	63	65.652	
184	62	66.211	
190	72	69.56	
	a =	0.5583	-36.50955 =b
		0.1131	19.66358
r=	0.818505	0.6699	3.492297
n=	14	24.358	12
t=	4.935	297.07	146.3537



$$t_{n-2} = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

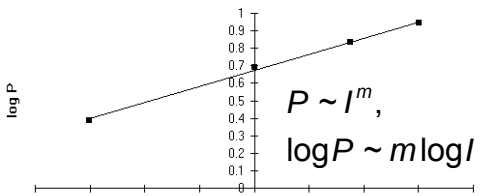
H_0: nincs korreláció
H_0: az alapsokaság korrelációs eh.-ja 0

$|t| = 4.935 > t_{12,\text{krit}(0,05)} = 2.179 \Rightarrow H_0$–t elvetjük (p<0.05)

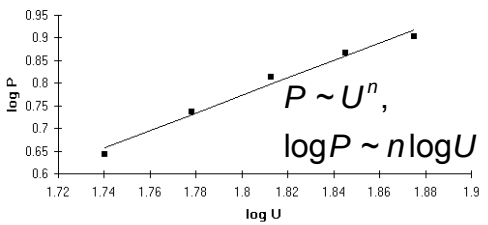
$|t| = 4.935 > t_{12,\text{krit}(0,01)} = 3.055 \Rightarrow H_0$–t elvetjük (p<0.01)

Példa: **Hatványfüggvényes regresszió visszavezetése lineáris regresszióra.** Röntgencső sugárteljesítményének mérése

	A	B	C	D	E
1	D1	97-10-16			
2					
3	I	P	logI	logP	
4	mA	r.e.			
5	0	0			
6	0.5	2.45	-0.30103	0.389166	0.396834
7	1	4.9	0	0.690196	0.673612
8	1.5	6.8	0.176091	0.832509	0.835517
9	2	8.8	0.30103	0.944483	0.950391
10					
11					
12	m=	0.919438	0.673612		
13		0.030392	0.007001		
14		0.99782	0.013744		
15		915.2208	2		
16		0.172873	0.000378		
17	r=	0.998909			
18					
19	U	P	logU	logP	
20	kV	r.e.			
21	55	4.4	1.740363	0.643453	0.656607
22	60	5.45	1.778151	0.736397	0.729975
23	65	6.5	1.812913	0.812913	0.797468
24	70	7.35	1.845098	0.866287	0.859957
25	75	8	1.875061	0.90309	0.918132
26					
27					
28	n=	1.941564	-2.72242		
29		0.145415	0.263339		
30		0.98345	0.015483		
31		178.2719	3		
32		0.042735	0.000719		
33	r=	0.991691			



$P \sim I^m$,
$\log P \sim m \log I$



$P \sim U^n$,
$\log P \sim n \log U$