Corelaţii si regresii

In domeniul medical se intalnesc numeroase elemente ce au o tendinta naturala de a se modifica impreuna.

Ex.:

- exercitiile fizice cresc frecventa cardiaca,
- oamenii inalti au masa mai mare,
- etc

Corelatia si regresia pun in evidenta relatiile ce exista intre doua serii de observatii considerate simultan.

De obicei aceste serii de obtin prin masurarea a doua caracteristici cantitative (variabile) pentru acelasi esantion.

Daca ne intereseaza doar existenta unei **legaturi** intre cele doua variabile, se calculeaza **coeficientul de corelatie**.

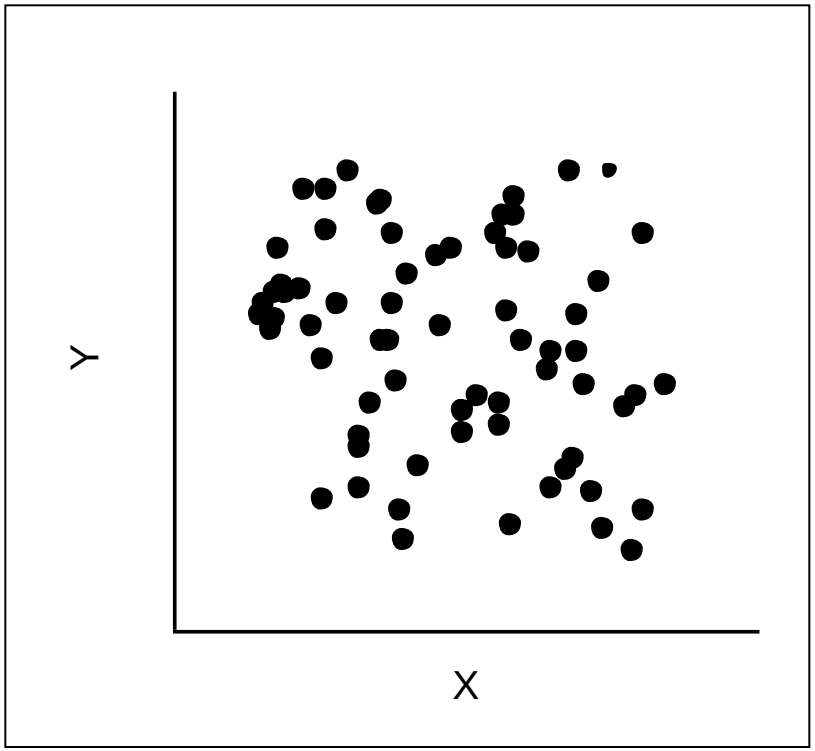
Un coeficient de corelatie mare indica o legatura puternica.

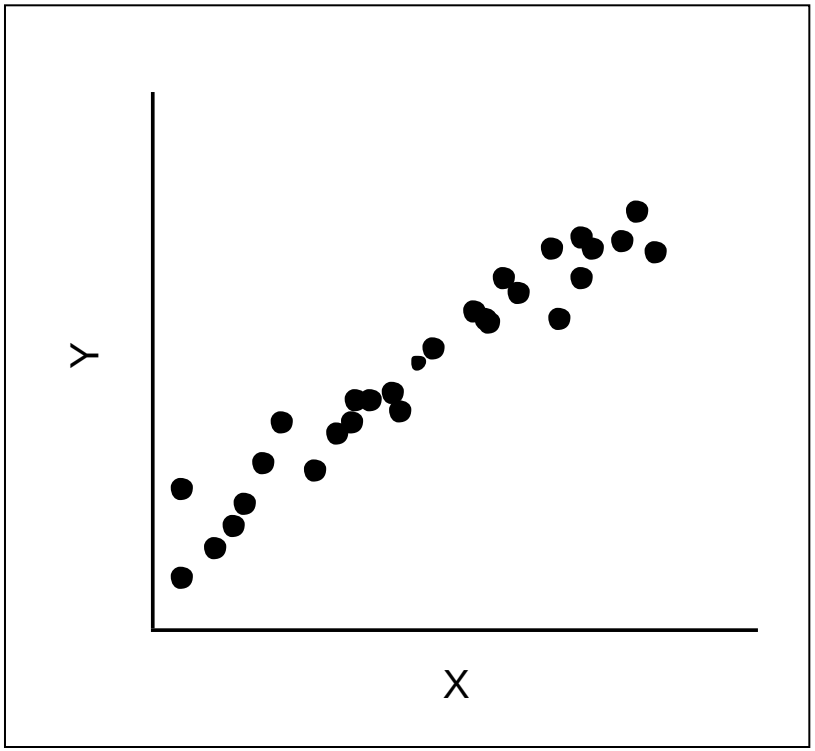
Daca ne intereseaza daca o variabila **depinde** de cealalta, si in ce fel, se determina **functia de regresie**.

Cele doua variabile sunt numite: variabila independenta si variabila dependenta.

Ce este corelaţia?

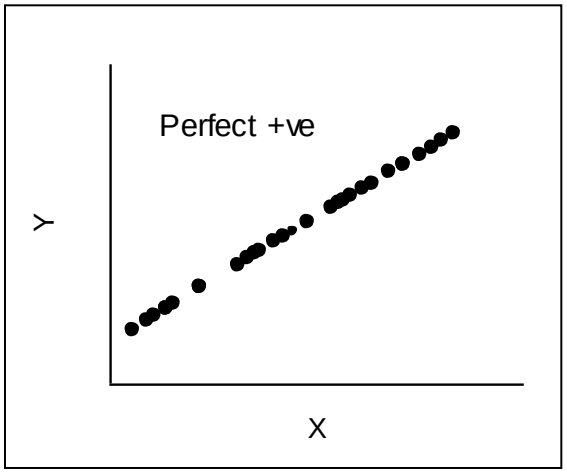
Corelatia (asocierea) dintre doua variabile se poate vizualiza cu ajutorul unei diagrame de dispersie



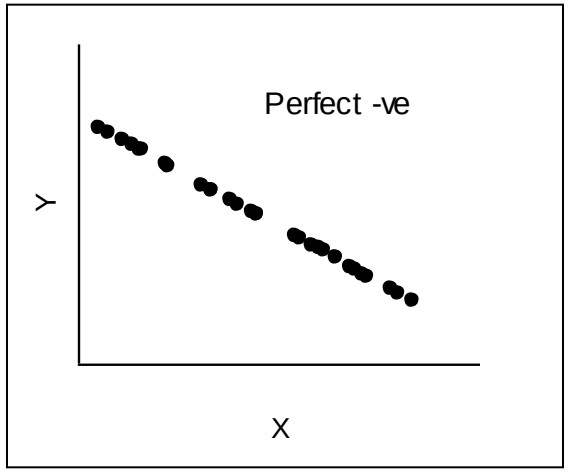


Masuratorile sunt pereche! Fluctuatiile celor doua variabile se “coreleaza” sufficient de bine pentru a exclude asocierea aleatoare. Totusi, corelarea statistica nu ne indica nici o cauzalitate.

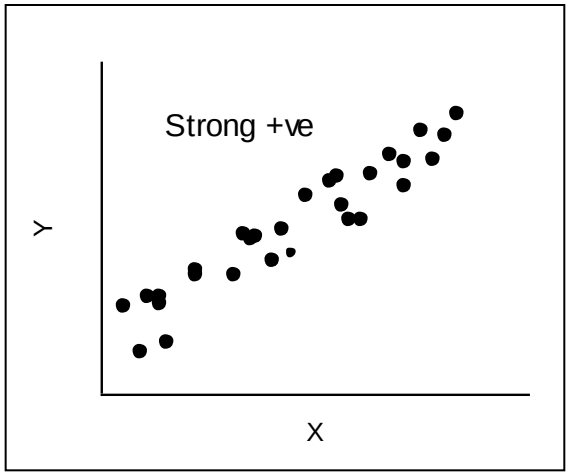
Tipuri de Corelaţie



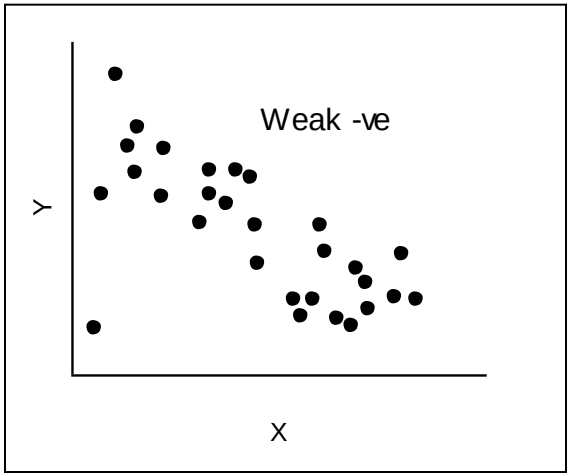
Corelatie perfect pozitiva



Corelatie perfect negativa

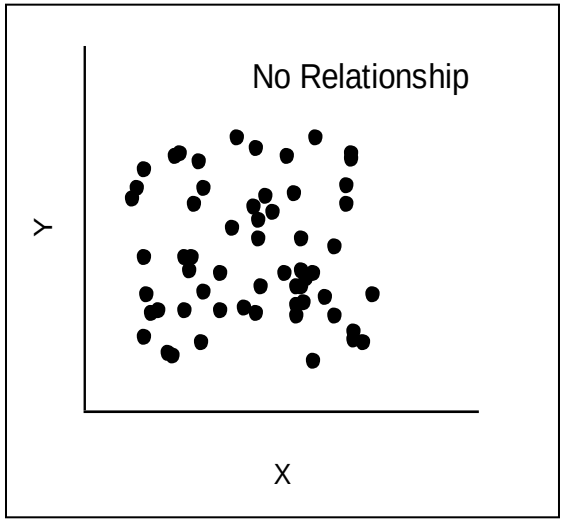


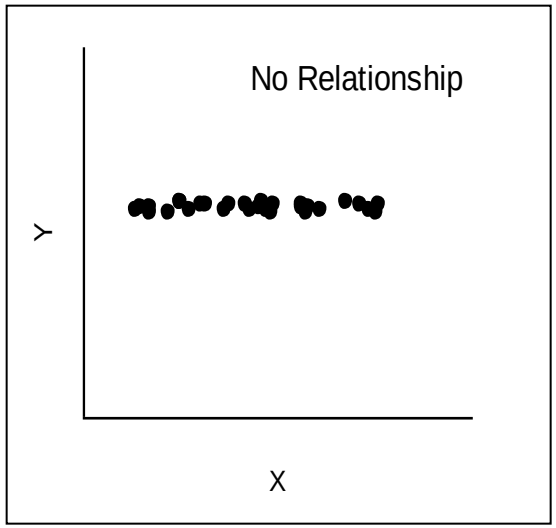
Corelatie puternic pozitiva

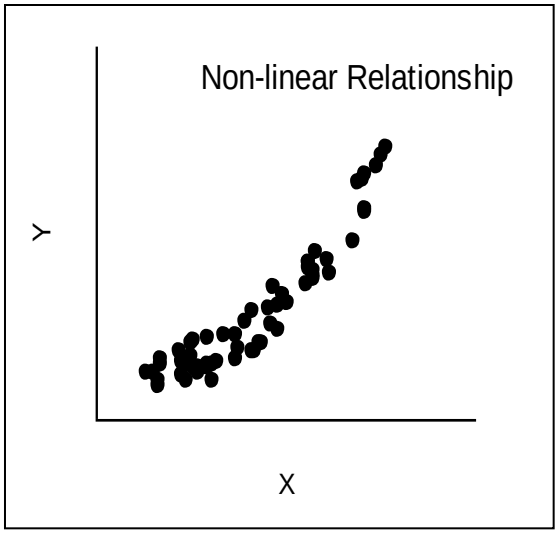


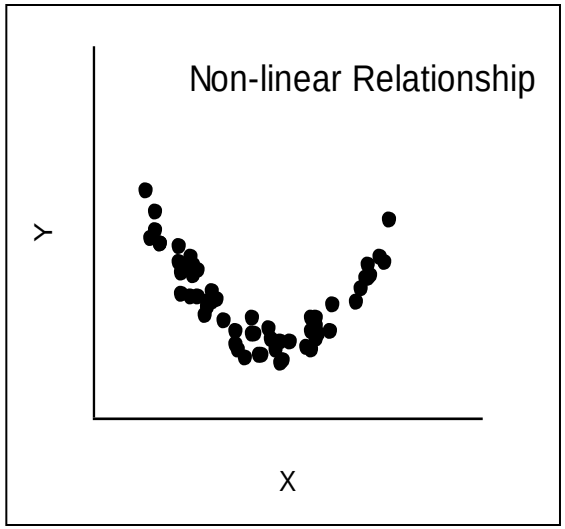
Corelatie slab negativa

Tipuri de Corelaţie









Corelatia are 3 caracteristici importante:

- **Directia:**
  - pozitiva (+)
  - negativa (-)
- **Forma:**
  - liniara
  - neliniara
- **Gradul de asociere**
  - intre -1 si +1
  - valoarea absoluta semnifica puterea asocierii

Coeficientii de corelaţie

- reprezinta o măsură a corelaţiei
- sunt adimensionali

- au valori între -1 şi +1
  - -1 $\Rightarrow$ corelaţie perfect negativă
  - +1 $\Rightarrow$ corelaţie perfect pozitivă
  - 0 $\Rightarrow$ nu există corelaţie (asociere aleatoare)

- Tipuri de coeficienţi
  - Coeficient Pearson $\mathbf{r_{xy}}$
  - Coeficient Spearman $\mathbf{r_s}$ (a ordinului)

Coeficientul de corelaţie Pearson

$$-1 < r_{xy} < +1$$

Observatie: Cu cat valoarea coeficientului de corelatie Pearson se apropie de de 1 (in valoare absoluta), cu atat "intensitatea" relatiei liniare dintre cele 2 variabile va fi mai mare!

Limite ale coeficientului Pearson:

Calculul se poate face numai pentru date scalate pe un interval.
Este un coeficient parametric, deci variabilele trebuie să fie normal distribuite.
Relaţia dintre cele două variable trebuie să fie liniară si sa aibă o “direcţie”.

Coeficientul de corelaţie Pearson

$$r_{xy} = \frac{\frac{1}{n}\sum (x - \bar{x})(y - \bar{y})}{s_x s_y}$$

n = mărimea eşantionului

x = valorile individuale ale variabilei x

y = valorile individuale ale variabilei y

$\bar{x}$ = media aritmetică a tuturor valorilor x

$\bar{y}$ = media aritmetică a tuturor valorilor y

s_x = deviaţia standard a tuturor valorilor x

s_y = deviaţia standard a tuturor valorilor y

$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y} = \frac{s_{xy}}{s_x \cdot s_y}$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

deviatiile standard pentru variabilele x si y

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}$$

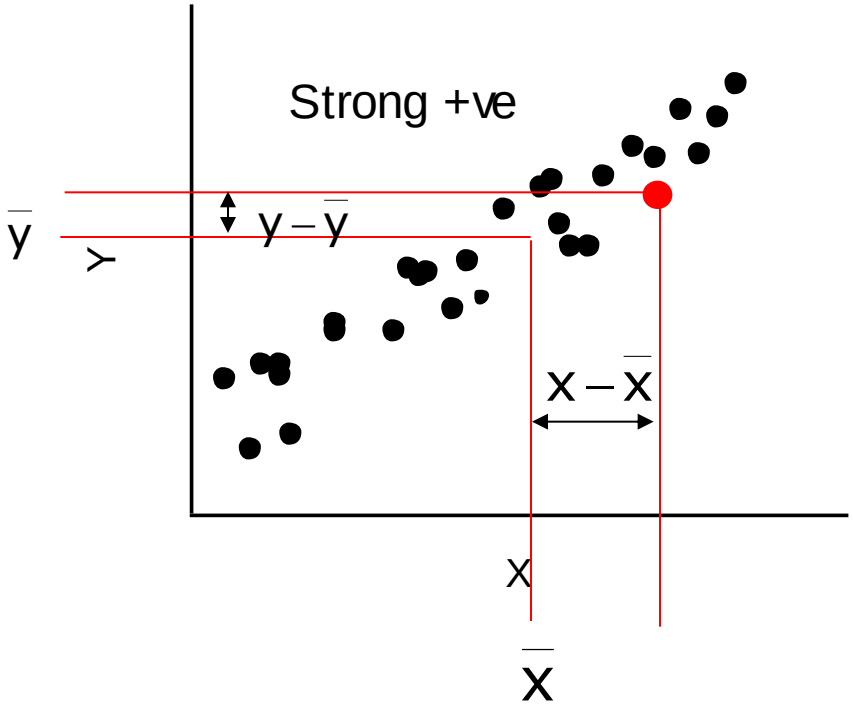
S_{xy}	**Covarianta**
$s_x \cdot s_y$	**Varianta totala**

$$r_{xy} = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}}$$

Covarianţa

$$s_{xy} = \frac{1}{n}\sum (x - \bar{x})(y - \bar{y})$$

Covarianta (*variabilitate pereche*) este independenta de marimea esantioanelor



$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Covarianţa (s_{xy})- exemplu de calcul

$$s_{xy} = 1/n \sum (x - \bar{x})(y - \bar{y})$$

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
5	4	2	1	2
3	1	0	-2	0
1	2	-2	-1	2
2	5	-1	2	-2
4	3	1	0	0
3	3	$1/n \sum (x - \bar{x})(y - \bar{y})$		$= 2/5 = 0{,}4$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}$$

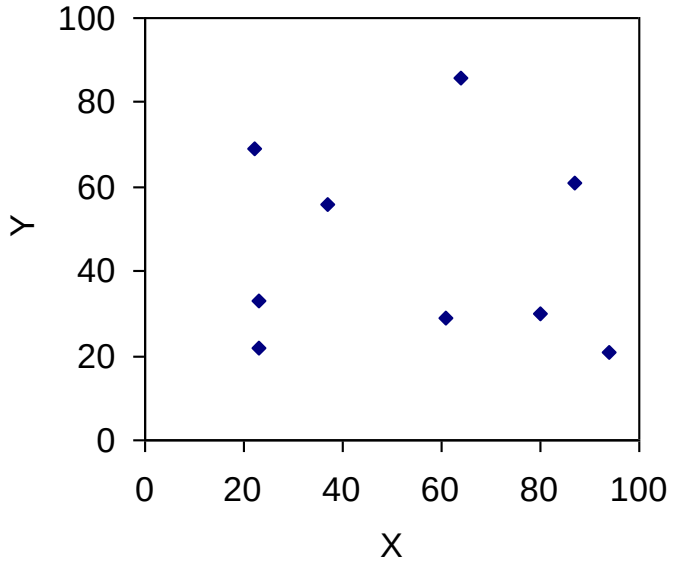
$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y} = \frac{S_{xy}}{s_x \cdot s_y}$$

$s_x \cdot s_y = 1{,}41 \cdot 1{,}41 \approx 2{,}0$

Varianţa totală $\rightarrow s_x \cdot s_y$

Coeficientul Pearson r_{xy} - exemplu de calcul

x	80	61	23	94	87	37	64	22	23
y	30	29	33	21	61	56	86	69	22



Coeficientul Pearson r_{xy} - exemplu de calcul

$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y}$$

	x	y
	80	30
	61	29
	23	33
	94	21
	87	61
	37	56
	64	86
	22	69
	23	22
Media	54,56	45,22
Dev St	27,38	22,02

Coeficientul Pearson r_{xy} - exemplu de calcul

$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y}$$

	x	y	$(x - \bar{x})$	$(y - \bar{y})$
	80	30	25,44	-15,22
	61	29	6,44	-16,22
	23	33	-31,56	-12,22
	94	21	39,44	-24,22
	87	61	32,44	15,78
	37	56	-17,56	10,78
	64	86	9,44	40,78
	22	69	-32,56	23,78
	23	22	-31,56	-23,22
Media	54,56	45,22		
Stdev	27,38	22,02		

Coeficientul Pearson r_{xy} - exemplu de calcul

$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y}$$

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
	80	30	25,44	-15,22	-387,32
	61	29	6,44	-16,22	-104,54
	23	33	-31,56	-12,22	385,68
	94	21	39,44	-24,22	-955,43
	87	61	32,44	15,78	511,90
	37	56	-17,56	10,78	-189,21
	64	86	9,44	40,78	385,12
	22	69	-32,56	23,78	-774,10
	23	22	-31,56	-23,22	732,79
Media	54,56	45,22			
Stdev	27,38	22,02			

Coeficientul Pearson r_{xy} - exemplu de calcul

$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y}$$

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
	80	30	25.44	-15.22	-387,32
	61	29	6.44	-16.22	-104,54
	23	33	-31.56	-12.22	385,68
	94	21	39.44	-24.22	-955,43
	87	61	32.44	15.78	511,90
	37	56	-17.56	10.78	-189,21
	64	86	9.44	40.78	385,12
	22	69	-32.56	23.78	-774,10
	23	22	-31.56	-23.22	732,79
Media	54,56	45,22	$\sum (x-\bar{x})(y-\bar{y})$		-395,11
Stdev	27,38	22,02			

Coeficientul Pearson r_{xy} - exemplu de calcul

$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y}$$

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
	80	30	25.44	-15.22	-387.32
	61	29	6.44	-16.22	-104.54
	23	33	-31.56	-12.22	385.68
	94	21	39.44	-24.22	-955.43
	87	61	32.44	15.78	511.90
	37	56	-17.56	10.78	-189.21
	64	86	9.44	40.78	385.12
	22	69	-32.56	23.78	-774.10
	23	22	-31.56	-23.22	732.79
Mean	54.56	45.22		$\sum (x-\bar{x})(y-\bar{y})$	-395,11
Stdev	27.38	22.02			
				$1/n \sum (x-\bar{x})(y-\bar{y})$	-43,90

Coeficientul Pearson r_{xy} - exemplu de calcul

$$r_{xy} = \frac{\frac{1}{n} \cdot \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y}$$

$s_{xy} = 43,90$

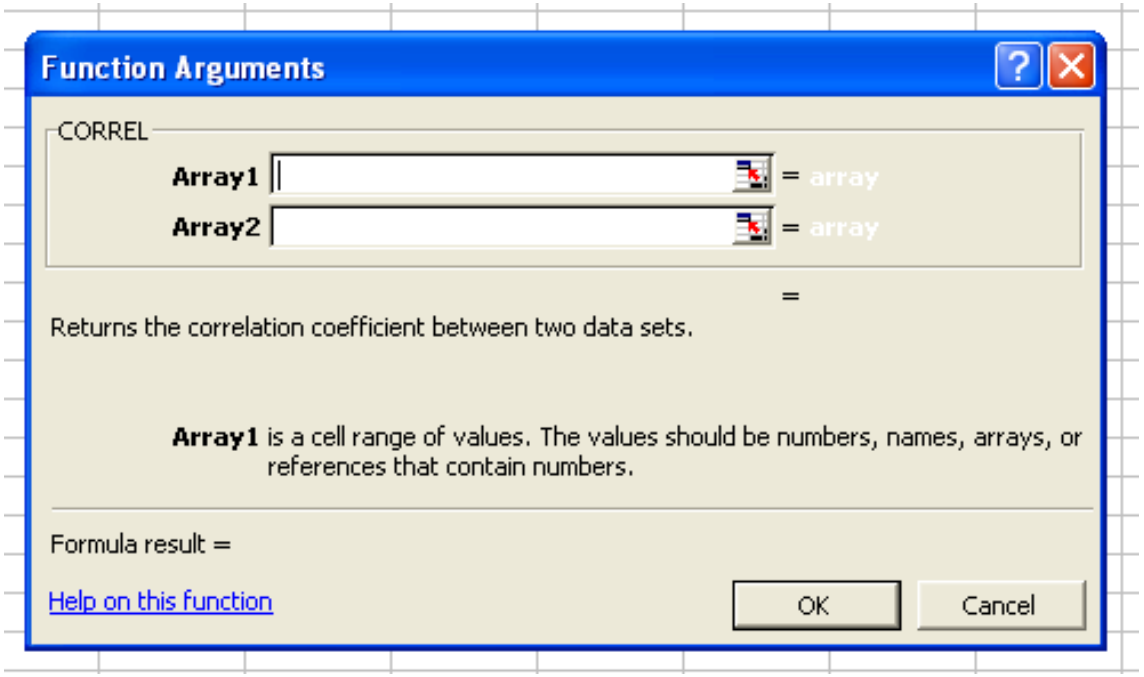
$s_x = 27,38$

$s_y = 22,02$

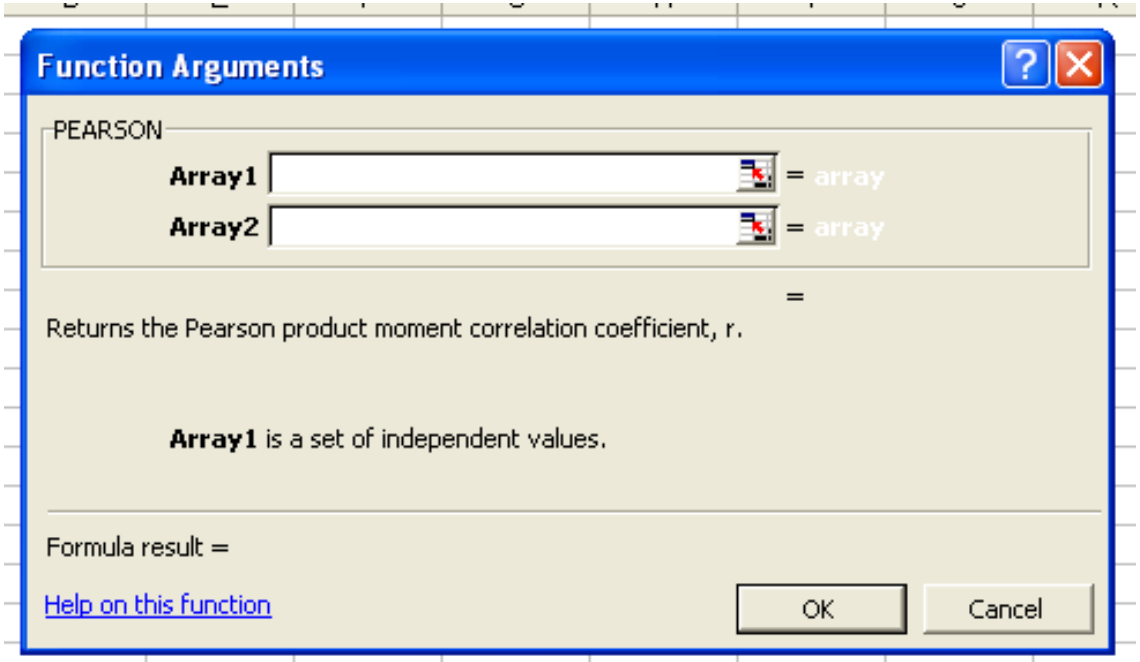
$$r_{xy} = \frac{1/n \sum (x - \bar{x})(y - \bar{y})}{s_x s_y} = \frac{-43,90}{27,38 \cdot 22,02} = -0,07$$

r_{xy} foarte mic → variabilele nu sunt corelate!

CORELL



calculeaza coeficientul de corelatie Pearson dintre 2 seturi de date



PEARSON

Regulile lui Colton (enuntate in 1974):

Un coeficient de corelatie de la -0,25 la 0,25 inseamna o corelatie foarte slaba sau nula.

Un coeficient de corelatie de la 0,25 la 0,50 (sau de la -0,25 la -0,50) inseamna o corelatie slaba (grad de asociere acceptabil)

Un coeficient de corelatie de la 0,50 la 0,75 (sau de la -0,50 la -0,75) inseamna o corelatie moderata spre buna

Un coeficient de corelatie mai mare de 0,75 (sau mai mic de -0,75) inseamna o corelatie puternica (grad de asociere foarte bun).

-0,25 < r < 0,25	Fara corelatie
-0,5 < r < -0,25 sau 0,25 < r < 0,5	Corelatie slaba
-0,75 < r < -0,5 sau 0,5 < r < 0,75	Corelatie moderata
r < -0,75 sau r > 0,75	Corelatie puternica

Pentru o interpretare corecta, coeficientul de corelatie trebuie sa fie insotit de un **test de semnificatie** (se determina valoarea nivelului de semnificatie α).

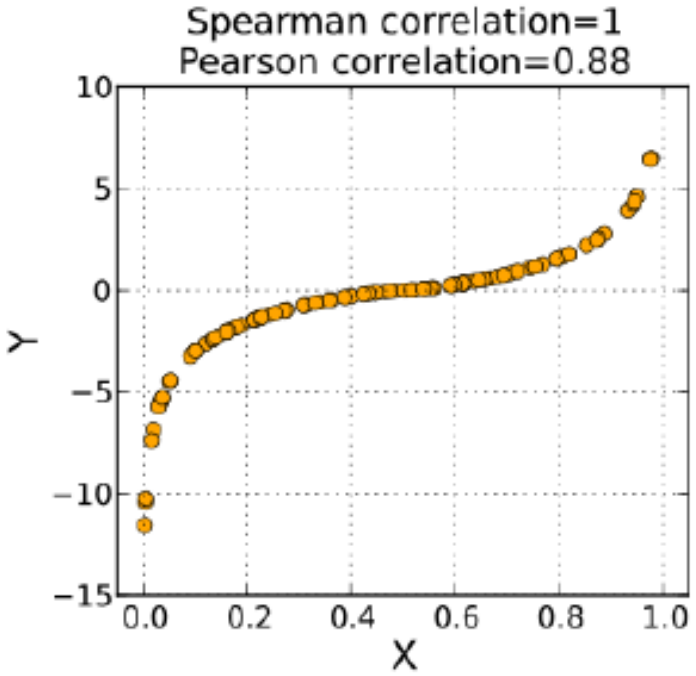
Interpretarea rezultatelor este sintetizata in tabel:

Coeficientul de corelatie	$\alpha > 0,05$	**$\alpha < 0,05$**
-0,25 < r < 0,25	Corelatie slaba sau nula	**Corelatie slaba sau nula**
-0,5 < r < -0,25 0,25 < r < 0,5	Nu are semnificatie statistica	**Grad de asociere acceptabil**
-0,75 < r < -0,5 0,5 < r < 0,75	Nu are semnificatie statistica	**Corelatie moderata spre buna**
r < -0,75 r > 0,75	Nu are semnificatie statistica	**Corelatie foarte buna**
r < -1 r > 1	Eroare	Eroare

Coeficientul Spearman r_s (de corelaţie a ordinului)

Coeficientul Spearman r_s rezolvă unele limite ale coeficientului Pearson r_{xy}

- Este non-parametric – nu se face nici o presupunere asupra normalităţii variabilelor.
- Relaţia dintre cele două variabile nu trebuie să fie liniară.
- Relaţia dintre cele două variabile trebuie să aibă o “direcţie”.
- Nu necesită date scalate pe un interval.



r_s = 1 inseamna ca relatia dintre cele doua variabile este monotona

Coeficientul Spearman r_s (de corelaţie a ordinului)

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

n = mărimea eşantionului

d = diferenţa între **ordinele** fiecărei perechi de valori ($d = r_x - r_y$)

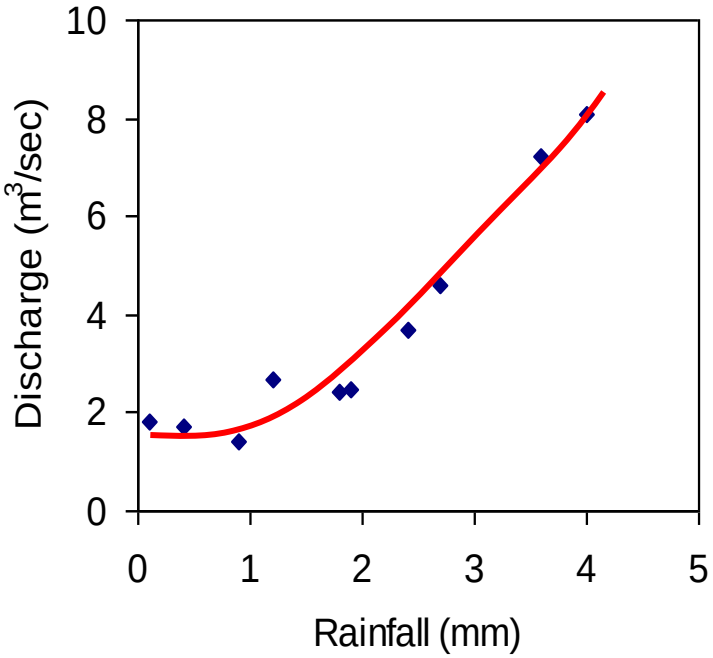
Limite ale coeficientului Spearman:

Sunt necesare date interval sau date ordinale.
Este necesar sa nu fie multe ordine “prea apropiate” în fiecare esantion
Se presupune că relaţia are o direcţie.
Coeficientul Spearman e mai puţin puternic decât coeficientul Pearson (~90%)

Coeficientul Spearman r_s (exemplu de calcul)

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

x	1.2	1.8	4.0	3.6	1.9	2.4	2.7	0.4	0.1	0.9
y	2.7	2.4	8.1	7.2	2.5	3.7	4.6	1.7	1.8	1.4



Coeficientul Spearman r_s (exemplu de calcul)

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

x	y	Ordinul lui x	Ordinul lui y	d (r_x-r_y)	d^2
1.2	2.7	4			
1.8	2.4	5			
4.0	8.1	10			
3.6	7.2	9			
1.9	2.5	6			
2.4	3.7	7			
2.7	4.6	8			
0.4	1.7	2			
0.1	1.8	1			
0.9	1.4	3			

Coeficientul Spearman r_s (exemplu de calcul)

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

x	y	Ordinul lui x	Ordinul lui y	d (r_x-r_y)	d^2
1.2	2.7	4	6		
1.8	2.4	5	4		
4.0	8.1	10	10		
3.6	7.2	9	9		
1.9	2.5	6	5		
2.4	3.7	7	7		
2.7	4.6	8	8		
0.4	1.7	2	2		
0.1	1.8	1	3		
0.9	1.4	3	1		

Coeficientul Spearman r_s (exemplu de calcul)

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

x	y	Ordinul lui x	Ordinul lui y	d (r_x-r_y)	d^2
1.2	2.7	4	6	-2	
1.8	2.4	5	4	+1	
4.0	8.1	10	10	0	
3.6	7.2	9	9	0	
1.9	2.5	6	5	+1	
2.4	3.7	7	7	0	
2.7	4.6	8	8	0	
0.4	1.7	2	2	0	
0.1	1.8	1	3	-2	
0.9	1.4	3	1	+2	

Coeficientul Spearman r_s (exemplu de calcul)

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

x	y	Ordinul lui x	Ordinul lui y	d (r_x-r_y)	d^2
1.2	2.7	4	6	-2	4
1.8	2.4	5	4	+1	1
4.0	8.1	10	10	0	0
3.6	7.2	9	9	0	0
1.9	2.5	6	5	+1	1
2.4	3.7	7	7	0	0
2.7	4.6	8	8	0	0
0.4	1.7	2	2	0	0
0.1	1.8	1	3	-2	4
0.9	1.4	3	1	+2	4

$\Sigma d^2 = 14$

Coeficientul Spearman r_s (exemplu de calcul)

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n}$$

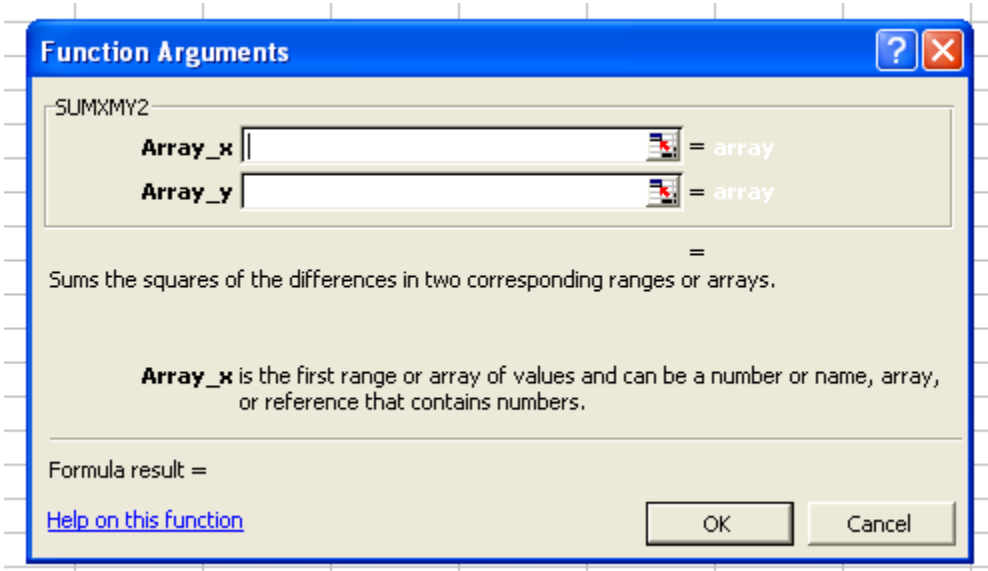
$\Sigma d^2 = 14$ n = 10

$$r_s = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6 \cdot 14}{10^3 - 10} = 1 - 0{,}085 = +0{,}915$$

r_s aproape de valoarea 1→ corelatie a ordinului foarte buna!

EXCEL: **SUMXMY2**

suma pătratele diferenţelor dintre 2 seturi de date



Testarea “semnificaţiei” coeficientului de corelaţie

Totdeauna trebuie verificat daca coeficientul de corelatie are semnificatie statistica sau nu!!!

Ipoteza nulului:

H_0: $\rho = 0$ (nu exista corelatie!)

Ipoteza alternativă (una din variantele):

H_1: $\rho > 0$ (corelaţie pozitivă)

H_1: $\rho < 0$ (corelaţie negativă)

H_1: $\rho \neq 0$ (exista corelatie, dar nu suntem siguri de semnul corelaţiei)

ρ - coeficientul de *corelaţie al populaţiilor din care sunt extrase cele doua variabile (esantioane)*

Testarea semnificaţiei coeficientului de corelatie (testul-t)

- Se converteşte r in unităţi t (se determina t_{calc})
- Dacă n ≥ 10 se poate folosi pentru ambii coeficienţi (r_{xy} şi r_s)
- Dacă n < 10 se foloseşte numai pentru r_{xy}

$$t_{calc} = r\sqrt{\left(\frac{n-2}{1-r^2}\right)}$$

- Se determina t_{crit} corespunzator gradului de libertate (df = n - 2) si nivelului de semnificatie (α = 0,05)
- **Coeficientul de corelatie are semnificatie statistica** ($\rho \neq 0$) **daca $t_{calc} > t_{crit}$**

Alta varianta:

- ***Coeficientul de corelatie are semnificatie statistica daca*** *valoarea nivelului de semnificatie α_{calc} (dedusa in functie t_{calc} folosind functia Excel TDIST) este < 0,05)*

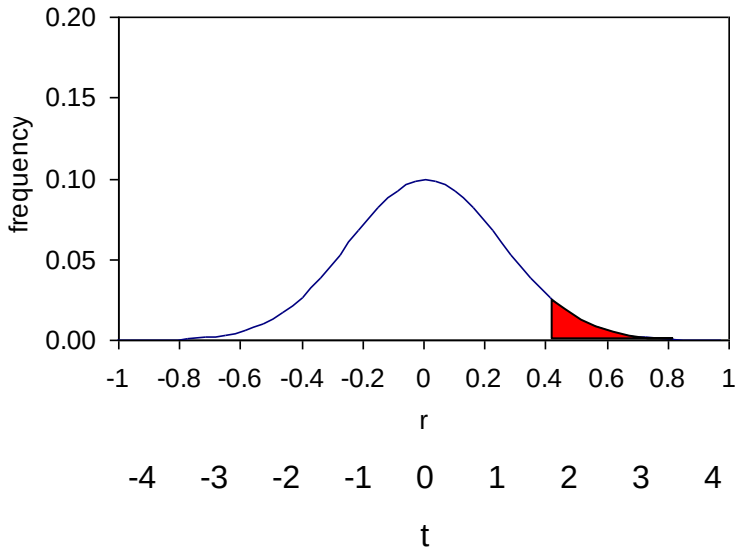
$$\alpha_{calc} < \alpha_{crit}$$

Testul pentru semnificaţia lui r poate fi "one-tailed" sau "two-tailed"!

Testele one-tailed se folosesc atunci cand se presupune tipul de corelatie (pozitiva sau negativa)

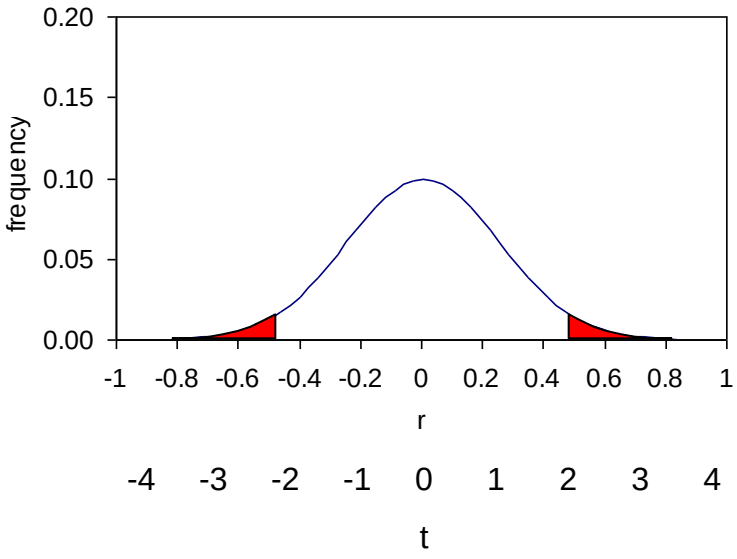
One-tailed

H_1: $\rho > 0$ or $\rho < 0$



Two-tailed

H_1: $\rho \neq 0$



Exemplu: testul-t pentru semnificaţia lui r

- r = -0,07; n = 9; df = 9 – 2 = 7
- H_0: $\rho = 0$
- H_1: $\rho \neq 0$

$$t_{calc} = r\sqrt{\left(\frac{n-2}{1-r^2}\right)} = -0.07\sqrt{\left(\frac{9-2}{1-(-0.07)^2}\right)} = -0{,}1857$$

- t_{crit} = 2,365 (α = 0,05; df = 7; two-tailed)
- $t_{calc} < t_{crit}$ … deci H_0 se accepta
  (ρ = 0: nu exista corelatie intre cele 2 seturi)

Concluzie: Coeficientul de corelatie nu are semnificatie statistica!

In locul testului t se poate folosii **un test echivalent**:

Se poate calcula **valoarea critica a coeficientului de corelatie** folosind formula :

$$r_{crit} = \frac{t_\alpha}{\sqrt{t_\alpha^2 + n - 2}}$$

Pentru a respinge H_0 trebuie ca: $r_{calc} \geq r_{crit}$

Pentru determinarea valorii critice a coeficientului de corelatie se poate folosi tabelul valorilor critice pentru r!

Coeficientul de corelatie Pearson

Tabelul valorilor critice pentru r_{xy}

Two tailed significance levels of the Pearson correlation coefficient.

	Significance Level				
df	**0.1000**	**0.0500**	**0.0250**	**0.0100**	**0.0050**
1	0.9877	0.9969	0.9992	0.9999	1.0000
2	0.9000	0.9500	0.9750	0.9900	0.9950
3	0.8054	0.8783	0.9237	0.9587	0.9740
4	0.7293	0.8114	0.8680	0.9172	0.9417
5	0.6694	0.7545	0.8166	0.8745	0.9056
6	0.6215	0.7067	0.7713	0.8343	0.8697
:	:	:	:	:	:
:	:	:	:	:	:
80	0.1829	0.2172	0.2475	0.2830	0.3072
90	0.1726	0.2050	0.2336	0.2673	0.2903
100	0.1638	0.1946	0.2219	0.2540	0.2759

For a sample size of n, df = n - 2

H_0 se respinge daca: $r_{xy\ calc} \geq r_{xy\ crit}$

Coeficientului de corelatie Spearman

Tabelul valorilor critice pentru r_s

Two tailed significance levels of the Spearman rank correlation coefficient.

	Significance Level			
N	**10%**	**5%**	**2%**	**1%**
4	**1.000**	**-**	**-**	**-**
5	**0.900**	**1.000**	**1.000**	**-**
6	**0.771**	**0.886**	**0.943**	**1.000**
7	**0.714**	**0.786**	**0.892**	**0.929**
8	**0.643**	**0.738**	**0.810**	**0.857**
9	**0.600**	**0.683**	**0.783**	**0.817**
10	**0.564**	**0.648**	**0.733**	**0.781**
11 or more	**Use a table for Pearson's r or the t-test**			

H_0 se respinge daca: $r_{s\ calc} \geq r_{s\ crit}$

Un **test rapid** (aproximativ) pentru testarea semnificatiei coeficientului de corelatie in cazul in care α = 0.05 este **$|r| > 2/\sqrt{n}$**

:

Sample Size	*Quick Rule*	*Quick* $r_{critical}$	*Actual* $r_{critical}$
$n = 25$	$\lvert r\rvert > \frac{2}{\sqrt{25}}$	.400	.396
$n = 50$	$\lvert r\rvert > \frac{2}{\sqrt{50}}$	.283	.279
$n = 100$	$\lvert r\rvert > \frac{2}{\sqrt{100}}$	.200	.197
$n = 200$	$\lvert r\rvert > \frac{2}{\sqrt{200}}$	.141	.139

Coeficientul de determinare (r^2)

- Ne spune cat din variaţia unei variabile este explicată prin variaţia celeilalte variabile.
- **Coeficientul de determinare este pătratul coeficientului Pearson r_{xy}**
- Coeficientul de determinare indica procentul din variatia totala a variabilei dependente care este explicata de variabila independenta

Exemple:

r = 0,60 ($r^2 = 0,6^2 = 0,36$)

36% din variaţia unei variabile este explicată prin variaţia celeilalte variabile.

Dacă r = 0,80 atunci variabilitatea variabilei independente explică 64% din variabilitatea variabilei dependente

P1. Intr-un studiu s-au calculat valorile coeficientului de corelatie intre greutatea si inaltimea pacientilor, rezultand o valoare de -0,87.

Nivelul de semnificatie obtinut este $\alpha_{calc} = 0,055$

Cum interpretam rezultatul?

P2. Care din urmatoarele afirmatii este corecta daca coeficientul de determinarea dintre 2 seturi de masuratori este 0,8?

a) valoarea unei masuratori creste cu 0,8 cand cealalta creste cu 1

b) 64% din observatii se gasesc pe dreapta de regresie.

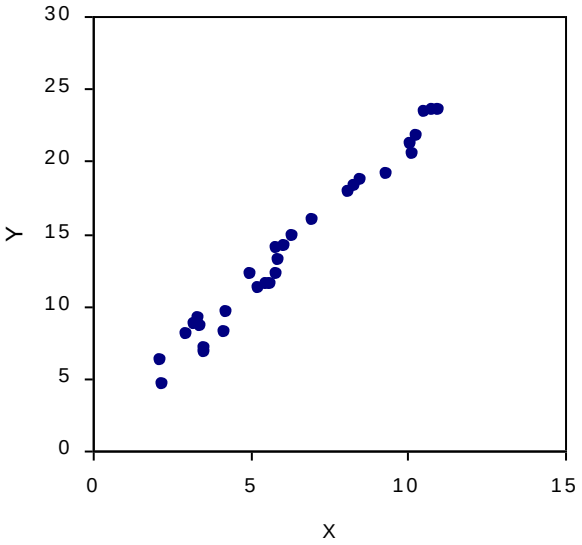
c) 80% din variatia unei masuratori este datorata celeilalte

d) 80% din observatii se gasesc pe dreapta de regresie.

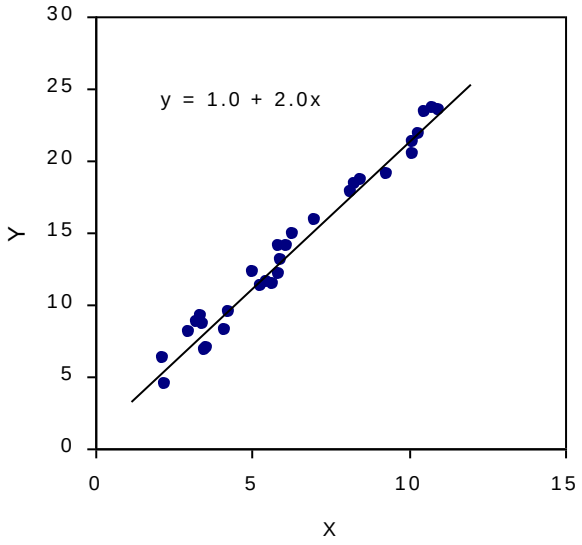
e) 64% din variatia unei masuratori este datorata celeilalte

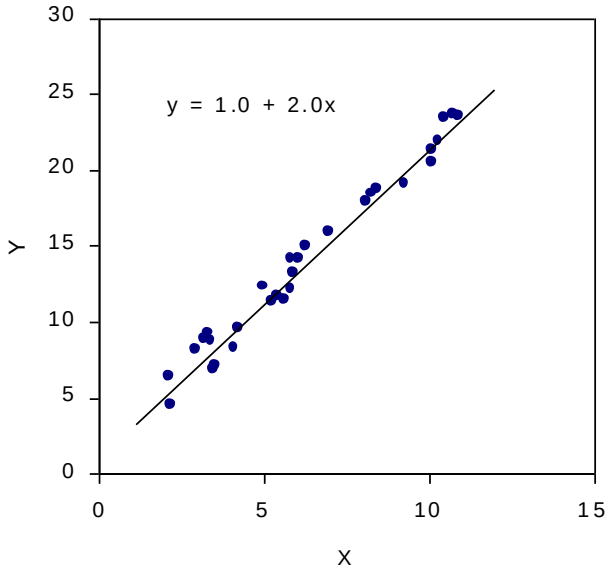
Regresia

- Corelaţia
  - ne spune că daca doua variabile sunt associate (correlate)
  - nu ne indica relatia dintre variabile



- Regresia
  - descrie relaţia dintre cele două variabile
  - ajuta la efectuarea de estimări
  - variabile dependente şi independente

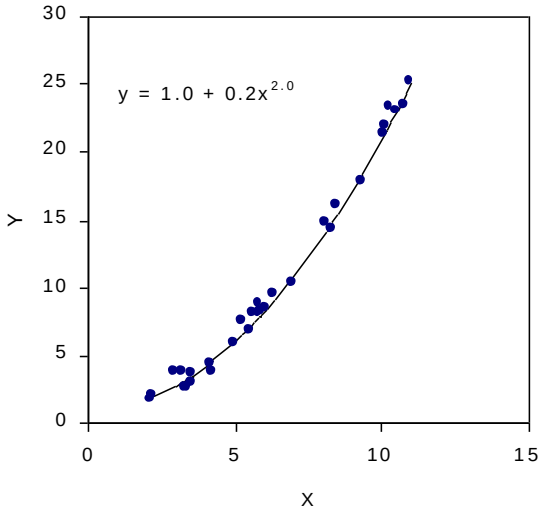




Regresie Liniară

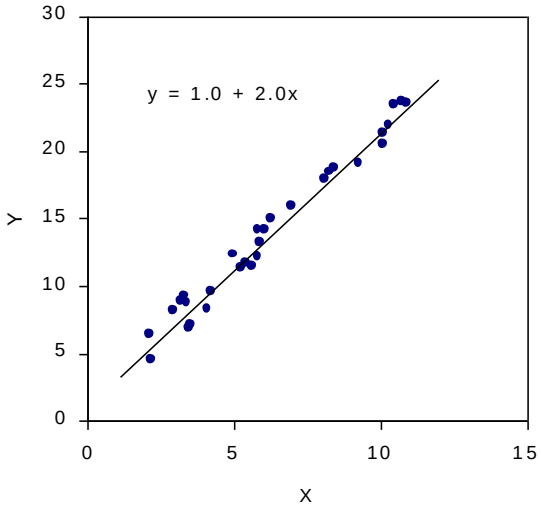
$$y = a + b \cdot x$$

Regresie neliniară



$$y = a + b \cdot x^{c}$$

Regresia Liniară



$$y = a + b \cdot x$$

x = variabilă independentă
y = variabilă dependentă

a, b, coeficienti de regresie

Scopul regresiei liniare este de a determina *dreapta de regresie* adica "linia dreapta care se potriveste cel mai bine datelor"

a → **intercept**: valoarea lui y când x este zero (intersecţia dreptei de regresie cu axa Oy).

b → **panta** dreptei de regresie: cantitatea cu care valoarea y se modifică în momentul în care modificăm valoarea lui x cu o unitate.

Coeficientii de regresie se pot calcula folosind:

- coeficientul de corelatie
- metoda celor mai mici patrate

I. Folosirea coeficientului de corelatie

$$b = r\frac{s_Y}{s_X}$$

$$a = \bar{y} - b\bar{x}$$

r - coeficientul de corelatie dintre variabilele X, Y

s_Y - deviatia standard a variabilei Y

s_X - deviatia standard a variabilei X

$\bar{y}$ - media variabilei Y

$\bar{x}$ - media variabilei X

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

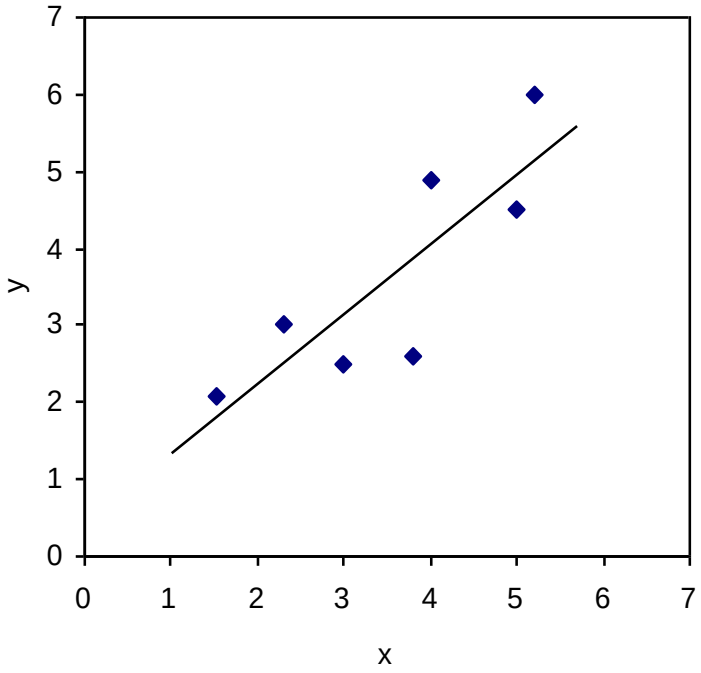
$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$$

Panta este influenţată de coeficientul de corelatie **r**, dar nu are aceeaşi semnificaţie!

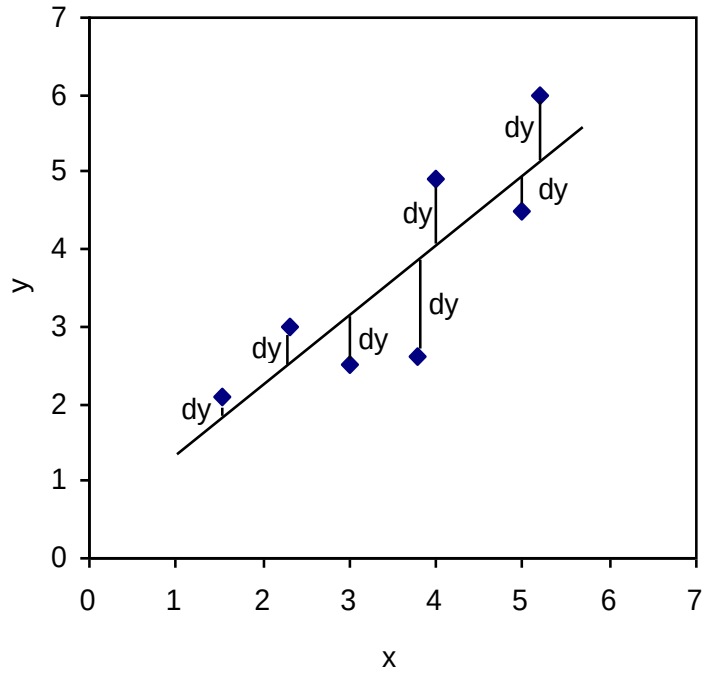
II. Metoda celor mai mici pătrate

- este utila in cazul in care cunoastem doar datele brute

- implica gasirea coeficientilor de regresie a si b astfel incat *suma patratelor reziduurilor sa fie minima.*



Obs.: Dreapta de regresie trebuie sa treaca cat mai aproape posibil de toate valorile observate!



$$y = a + b \cdot x$$

$$\sum dy = 0$$

$$\sum dy^2 = \text{minimum}$$

$$dy = y_i - y_i^{est}$$

Rezidurile *(dy)* sunt diferenţele dintre valorile actuale şi valorile estimate.

Reziduu = diferenţa dintre valoarea y observata şi valoarea y estimata prin introducerea lui x în ecuaţia de regresie.

Diferenta dintre valorile reale si cele estimate reprezinta *erorile de estimare* (valorile reziduale).

Media tuturor valorilor reziduale este zero!

Folosirea metodei celor mai mici pătrate implica rezolvarea sistemului:

$$n \cdot a + b \sum x = \sum y$$
$$a \sum x + b \sum x^2 = \sum (x \cdot y)$$

n - marimea esantioanelor (numarul perechilor din cele 2 seturi de date)

Σx - suma tuturor valorilor x,

Σy - suma tuturor valorilor y,

Σx^2 - suma tuturor patratelor valorilor x,

$\Sigma(x \cdot y)$ - suma tuturor produselor x·y.

Eroarea standard a estimarii $S_{y(x)}$

$$s_{y(x)} = s_y \cdot \sqrt{1 - r^2}$$

s_y - deviatia standard a variabilei y,

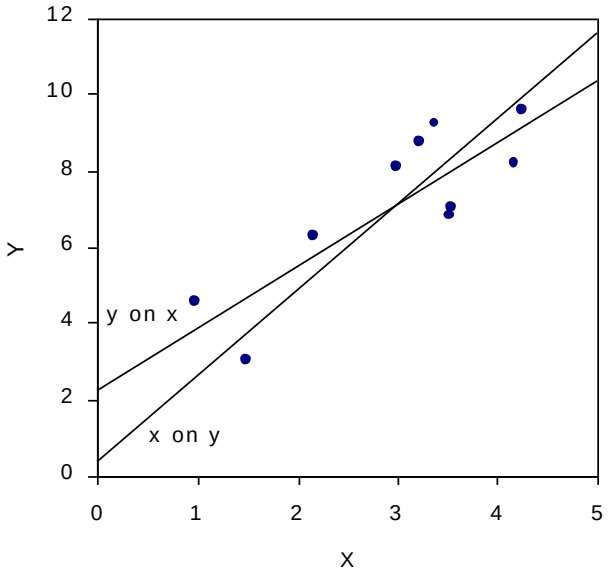
r - coeficientul de corelatie.

Cu cât coeficientul de corelatie este mai mare, cu atât eroarea de estimare va fi mai mică!

P3. Sa se determine ecuatia dreptei de regresie (prin metoda celor mai mici patrate) pentru seturile de date ce reprezinta valoarea calciului seric (y) si valoarea parathormonului (x)

$$n \cdot a + b\sum x = \sum y$$
$$a\sum x + b\sum x^2 = \sum (x \cdot y)$$

x	y	x^2	x·y
1,2	2,7		
1,8	2,4		
4	8,1		
3,6	7,2		
1,9	2,5		
2,4	3,7		
2,7	4,6		
0,4	1,7		
0,1	1,8		
0,9	1,4		
Σx = 19	Σy = 36,1	Σx^2 = 51,28	Σx·y= 94,05



Pentru orice set de date se pot trasa doua drepte de regresie care minimizeaza suma patratelor reziduurilor pe axele Ox respectiv Oy.

Daca valoarea y va fi estimata folosind valoarea x, atunci se foloseste dreapta care minimizeaza reziduurile pe axa Oy.

$$y = a + bx$$

$$n \cdot a + b\sum x = \sum y$$

$$a\sum x + b\sum x^2 = \sum (x \cdot y)$$

Daca valoarea x va fi estimata folosind valoarea y, atunci se foloseste dreapta care minimizeaza reziduurile pe axa Ox.

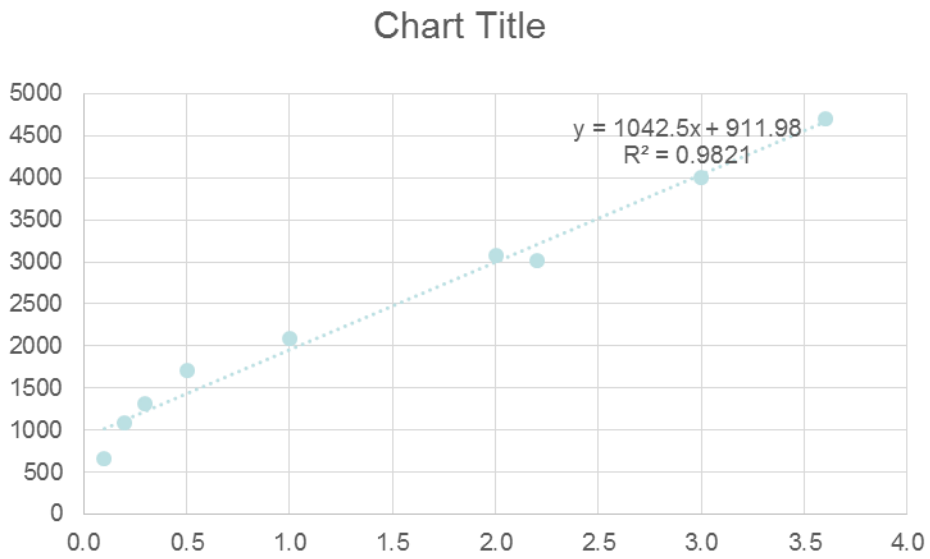
$$x = A + By$$

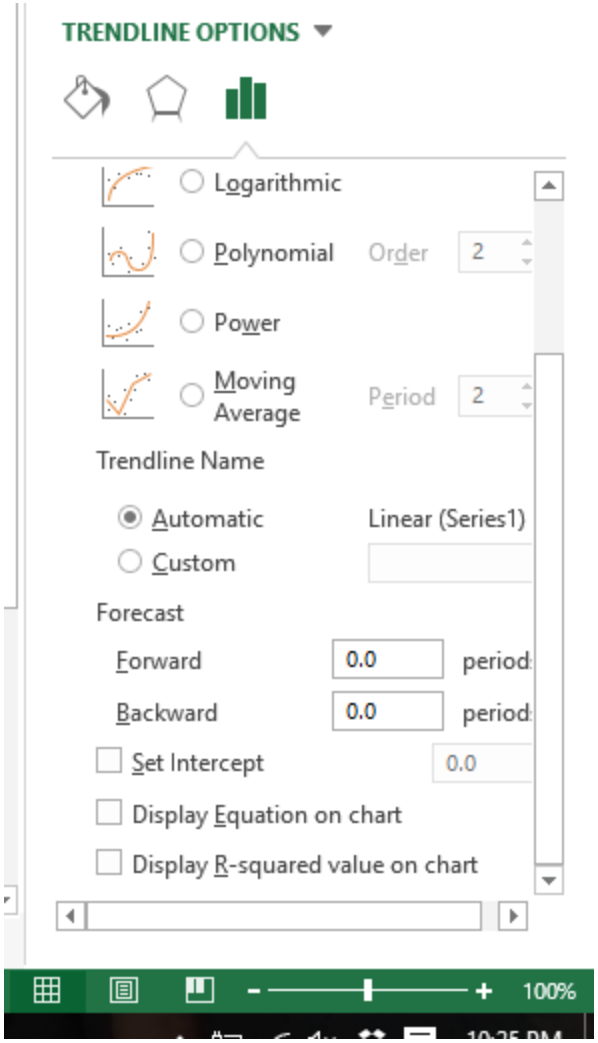
$$n \cdot A + B\sum y = x$$

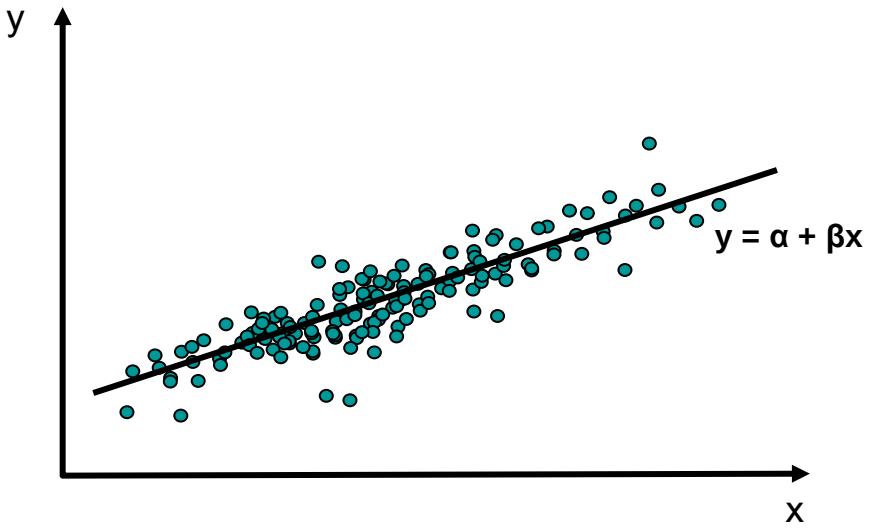
$$A\sum y + B\sum y^2 = \sum (x \cdot y)$$

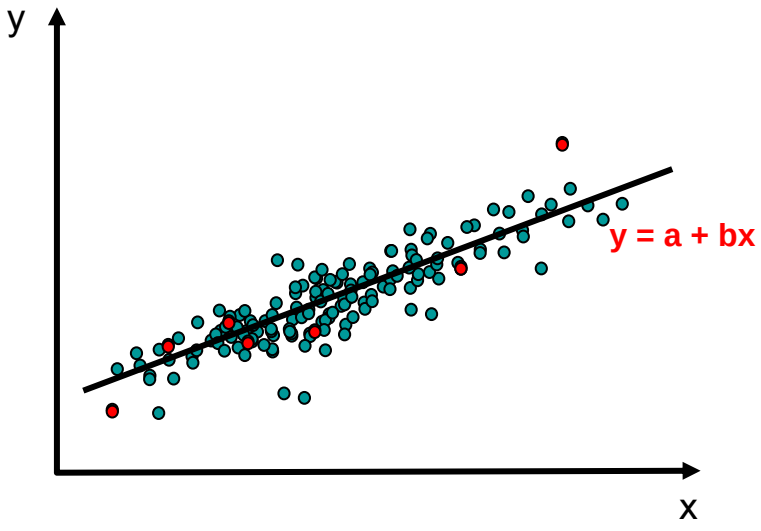
In Excel se poate determina ecuatie de regresie prin “fitarea” unui grafic tip Scatter:

Dupa ce ati realizat graficul (Scatter):
- “Click” pe unul din punctele din graphic pentru a selecta datele
- “Right-click” si alegeti “Add trend line”
- In “Tredline Options” selectati “Display Equation” si “Display R-squared value”

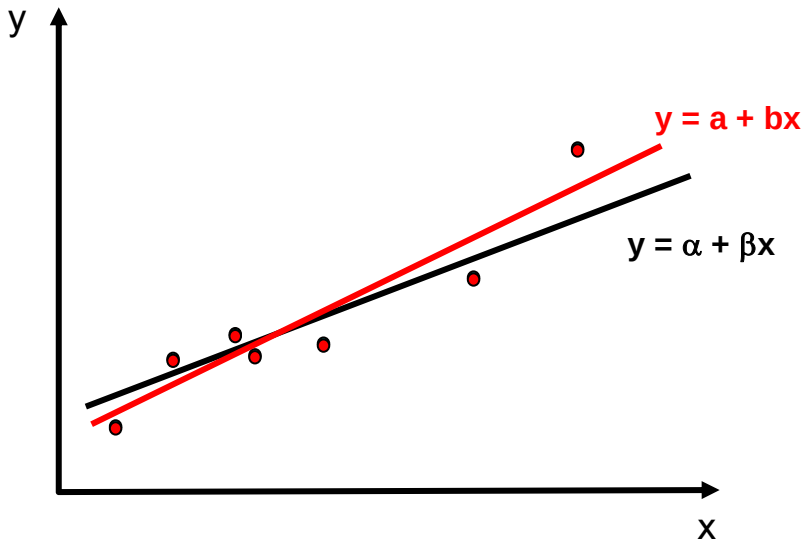


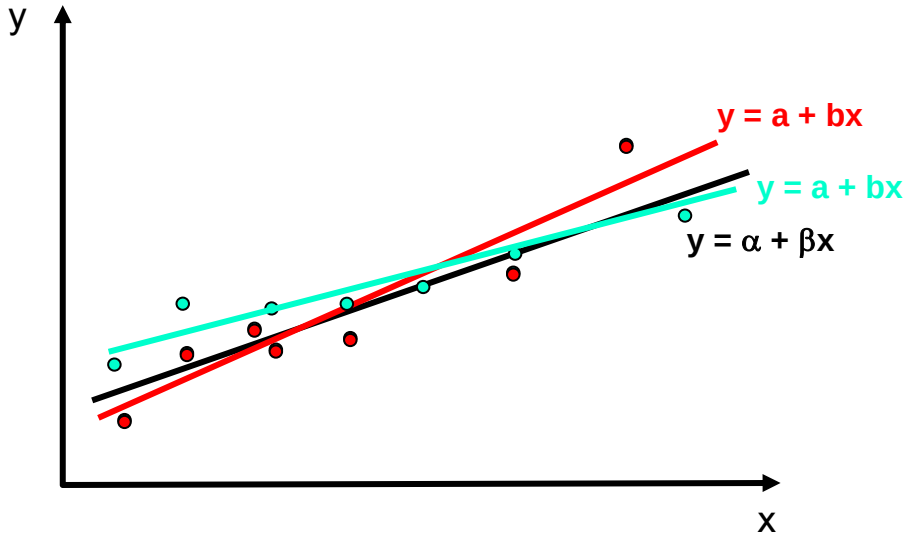






Dreapta de regresie a populaţiei nu este identica cu dreapta de regresie a esantionului!





Testarea semnificaţiei coeficienţilor de regresie

- Regresia prin metoda celor mai mici pătrate va da totdeauna o cea mai potrivită linie (best fit line) … dar trebuie **testat dacă această linie are semnificaţie statistică**.

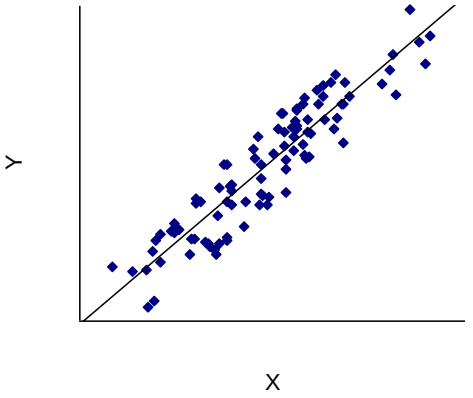
Când se fac deducţii statistice folosind regresia liniară se presupune că s-a eşantionat o populaţie ce are o relaţie liniară între x şi y, cu valori fixe (dar necunoscute) ale pantei (β) şi interceptului (α).

Valorile interceptului (a) şi pantei (b) calculate folosind esantioanele X şi Y ***estimează parametrii de regresie ai populaţiei*** *(*dau dreapta de regresie, care fitează întreaga populaţie)

Pentru a determina intervalul de incredere (pentru panta (β) şi interceptului (α) populatiei) si a testa semnificatia dreptei de regresie trebuie determinata **eroarea standard de predictive (s_{yx} sau s_{xy}):**

s_{yx} - daca y este estimat din valoarea lui x

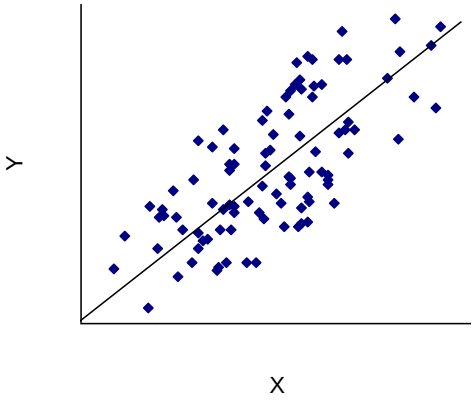
$$s_{yx} = \sqrt{\frac{\sum_i \left(y_i - y_i^{est}\right)^2}{n-2}}$$



$y_i \rightarrow$ valoarea variabilei dependente;

$y_i^{est} \rightarrow$ valoarea estimată folosind ecuaţia de regresie.

s_{xy} - daca x este estimat din valoarea lui y



Testul-t pentru pantă

- Testează dacă panta (b) are semnificaţie statistică

- Dacă în populaţiile mamă nu este nici o relaţie între x şi y, ne aşteptăm ca panta dreptei de regresie să fie zero

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

- Distribuţia valorii pantei (b) este o distribuţie de tip t

$$t_{calc} = b / S_b = b / \frac{S_{yx}}{\sqrt{\sum_{1}^{n} (x_i - \bar{x})^2}}$$

- Se transforma valoarea calculată pentru b în unităţi t (t_{calc}):

- Se determina valoarea critica (t_{crit}), definita de nivelul de semnificatie α si gradul de libertate (df = n - 2)

Testul-t pentru pantă

•Se compara valoarea calculata cu valoarea critica

•Se accepta una din ipotezele enuntate:

- daca $\mathbf{t_{calc} < t_{crit}}$, $\mathbf{H_0}$ **este acceptata** (panta dreptei de regresie nu are semniticatie statistica)

- daca $\mathbf{t_{calc} > t_{crit}}$, $\mathbf{H_0}$ **este respinsa** (panta dreptei de regresie are semniticatie statistica)

Cu cât valoarea lui t_{calc} este mai mare, cu atât e mai mică posibilitatea ca valoarea pantei să provină din eşantionarea aleatoare a unor variabile care nu sunt liniar relaţionate.

Testul-t pentru intercept

- Testează dacă interceptul (a) are semnificaţie statistică
- Se foloseşte testul-t: H_0: $\alpha = 0$

  H_1: $\alpha \neq 0$
- Se transforma valoarea calculată pentru **a** în unităţi t_{calc}

$$t_{calc} = a / S_a = a / S_{yx} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_1^n (x_i - \bar{x})^2}}$$

•Se determina valoarea critica (t_{crit}), definita de nivelul de semnificatie α si gradul de libertate (df = n - 2)

Se compara valoarea calculata cu valoarea critica

Se accepta una din ipotezele enuntate:

- daca **$t_{calc} < t_{crit}$, H_0 este acceptata**
- daca **$t_{calc} > t_{crit}$, H_0 este respinsa** (interceptul dreptei de regresie are semniticatie statistica)

Limitele regresiei liniare

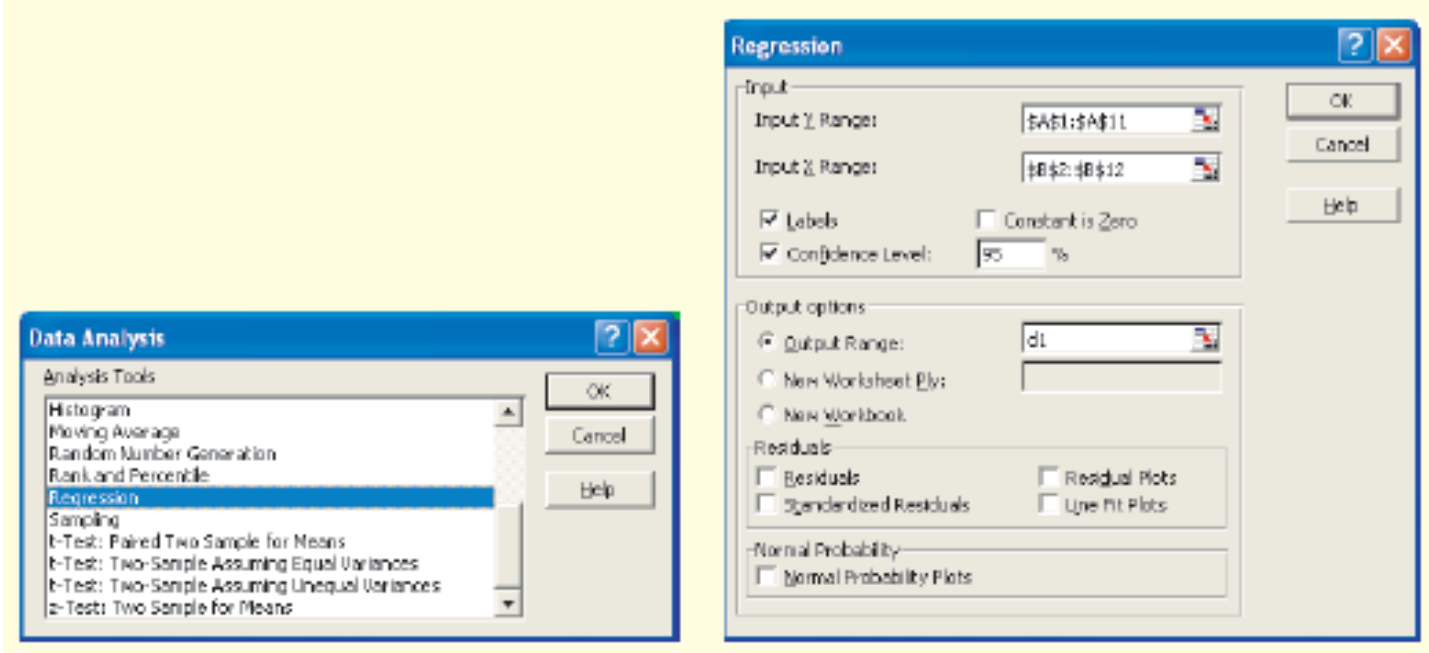
- Sunt necesare date scalate pe interval.
- Datele trebuie să fie aproximativ normal distribuite.
- Relaţia dintre variabile este presupusă liniara, dar uneori o fitare neliniară poate să dea rezultate mai bune pentru regresie.
- Precizia (acuratetea) măsurătorilor pentru variabila independentă se presupune a fi bună.
- Ecuaţia de regresie nu trebuie folosită pentru a prezice valori dincolo de limita datelor originale.

- Pentru orice valoare a lui x, valoarea corespunzătoare y este normal distribuită faţă de dreapta de regresie a populaţiei. (reziduurile regresiei trebuie să fie aproximativ normal distribuite, cu o medie egală cu zero).
- Varianţa lui y faţă de dreapta de regresie nu variază semnificativ pe domeniul lui x.
- Reziduurile nu trebuie să aibă vreo “tendinţă” (ex. panta dreaptei de regresie a reziduurilor trebuie să fie zero).
- Nu se poate rearanja ecuaţia de regresie din y funcţie de x, pentru a prezice x din y.

Analiza Varianţei (ANOVA)

- Testează dacă variaţia lui y ‘explicată’ prin ecuaţia de regresie este statistic semnificativă.
- Calculează raportul dintre varianţa explicată prin regresie şi varianţa reziduurilor – numit F
- Poate testa semnificaţia lui F
- Dacă F este mare, atunci proporţia varianţei în eţantion explicată prin ecuaţia de regresie este improbabil să provină din eşantionarea aleatoare a variabilelor populaţiei care nu au nici o relationare

Excel: Se selecteaza Regression din “Data Analysis”



Se obtin valorile pantei si interceptului precum si statistica lor

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.627790986
R Square	0.394121523
Adjusted R Square	0.318386713
Standard Error	14.00249438
Observations	10

Variable	*Coefficient*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	49.47712665	10.06646125	4.915047	0.001171	26.26381038	72.69044293
Study Hours	1.964083176	0.86097902	2.281221	0.051972	−0.021339288	3.94950564

$$r_{xy} = \frac{\frac{1}{n}\cdot\sum_i (x_i-\bar{x})\cdot(y_i-\bar{y})}{s_x\cdot s_y} = \frac{S_{xy}}{s_x\cdot s_y}$$

$$s_{yx} = \sqrt{\frac{\sum_i \left(y_i - y_i^{est}\right)^2}{n-2}}$$

$$t_{calc} = b/S_b = b \Big/ \frac{S_{yx}}{\sqrt{\sum_1^n (x_i-\bar{x})^2}}$$

Coeficientul de determinare r^2

r^2 = SSR/SST

Coeficientul de corelatie

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.991000941							
R Square	0.982082866							
Adjusted R Square	0.979523275							
Standard Error	196.9899126							
Observations	9							
ANOVA								
	df		SS	MS	F	Significance F		
Regression	1		**14889002.38**	14889002	383.6874767	2.25643E-07		
Residual	7		271635.1796	38805.03				
Total	8		**15160637.56**					
	Coefficients		Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	**911.9844282**		100.6521225	9.060757	4.08297E-05	673.9799784	1149.988878	673.9799784	1149.988878
X Variable 1	**1042.491484**		53.22108252	19.58794	2.25643E-07	916.6436218	1168.339347	916.6436218	1168.339347

Syx

SSR

SST

panta

S_b

t_{calc}

α_{calc}

Regresia Multiplă

- o variabilă dependentă
- Mai multe variabile independente
- $y = a + bx_1 + cx_2 + \ldots$

De reţinut!

▫ Evaluarea puterii asocierii dintre două variabile cantitative continue (normal distribuite) —› **corelaţie**

▫ Prezicerea unei variabile (Y) în funcţie de o altă variabilă (X) —› **regresie**