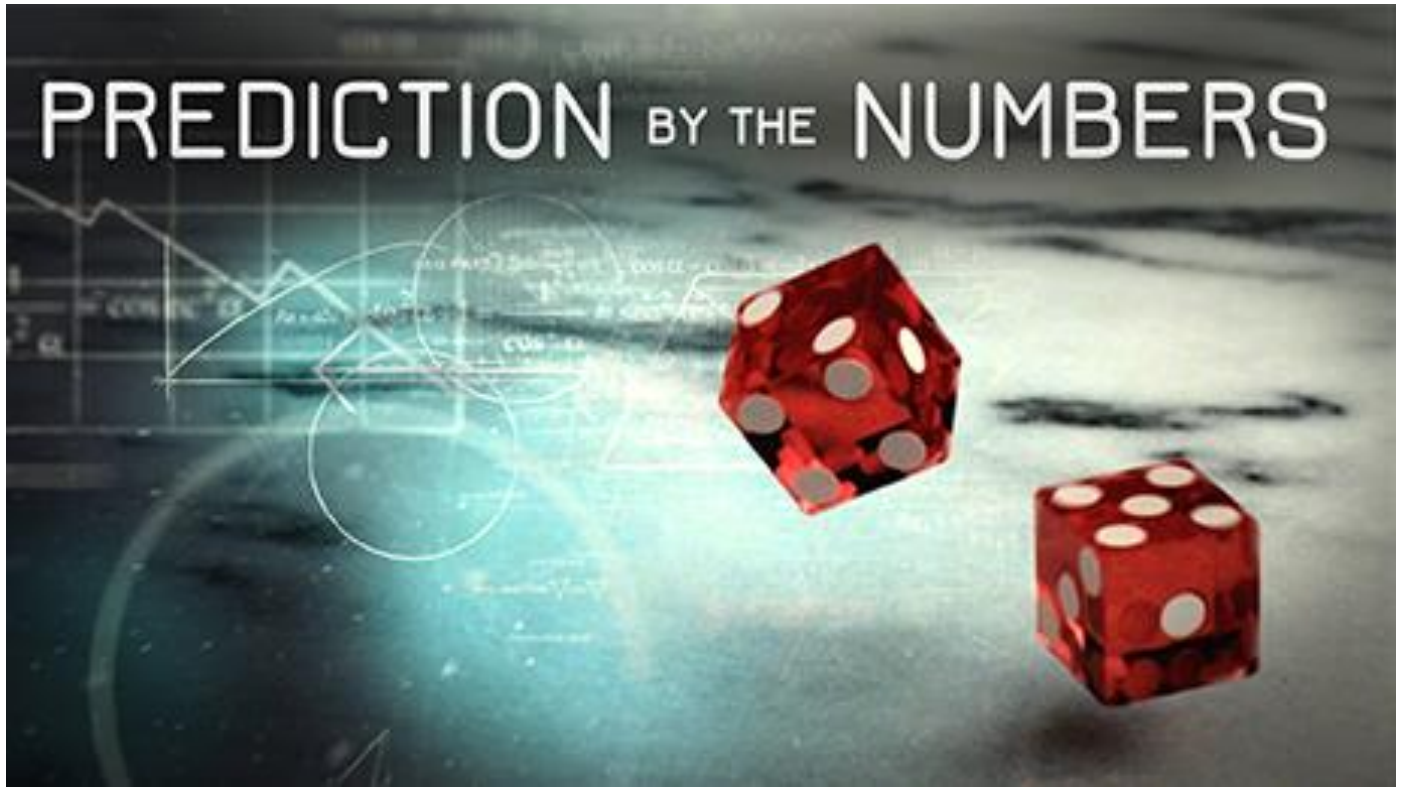


Autor: Bondor Cosmina-Ioana

Asociere şi predicţie

A ALWAYS

S SEEK

K KNOWLEDGE

Obiective

- Corelaţie
- Regresie liniară
- Exerciţii

Scenariu

- 36 de pacienţi cu artroplastie bilaterala totală a genunchiului
- A fost măsurată:
  - Greutatea
  - Densitatea minerală osoasă
- Scop: Există asociere între greutate şi densitatea minerală osoasă

Ishii Y, Noguchi H, Sato J, Ishii H, Todoroki K, Toyabe SI. Association between body weight and proximal tibial bone mineral density after bilateral total knee arthroplasty. Knee. 2017 Oct;24(5):1153-1159.

Corelaţie

- Relaţia dintre două caracteristici
- Cum este relaţia?
- Putem prezice un eveniment?
- Ce eroare putem să acceptăm?

Asociere

- Termeni similari: relaţie, dependenţă.
- Termenul **corelaţie** se utilizează numai în cazul a două variabile numerice sau ordinale

- Dacă Y poate fi calculat din X = asociere
- Ex. Y este de 3 ori mai mare decât X. Dacă îl ştim pe X, putem să-l calculăm pe Y.

Dependenţă

- Dacă X este cauza lui Y, atunci Y este **dependent** de X

Greutatea asociată cu densitatea minerală osoasă

- X = variabila independentă (greutatea)
- Y = variabila dependentă (densitatea minerală osoasă)

- Mai întâi X, apoi Y – demonstraţia cauzalităţii
- ! Ca să demonstrăm cauzalitatea avem nevoie de protocol de studiu specific, de mai multe studii, un studiu de corelaţie nu este suficient

Relaţia dintre două caracteristici

- Două variabile cantitative sau calitative odinale
  - relație liniară: coeficient de corelație Pearson
  - relație neliniară: coeficient de corelaţie Spearman
  - Coeficient de corelație interclase
- Două variabile calitative:
  - Risc relativ
  - Rata şansei
  - Corelația interobservator - Cohen Kappa coeficient de concordanță
- Dacă este implicată variabila de timp ca variabilă dependentă
  - Rata hazardului

Coeficient de corelaţie Pearson

Dacă X și Y sunt două variabile cantitative sau ordinale. Coeficientul de corelație Pearson:

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$$

Unde n - numărul de pacienți (dimensiunea eșantionului), $\overline{X}$,$\overline{Y}$-media aritmetic a variabilelor X respectiv Y

Coeficient de corelaţie Pearson

- Indică asocierea dintre X şi Y
- r=1 sau r=-1 corelaţie perfectă

- Este între -1 şi 1

$$r \in [-1,1]$$

Cu cât $|r|$ se apropie de 1 cu atât asocierea este mai puternică

Cu cât $|r|$ se apropie de 0 cu atât asocierea este mai slabă

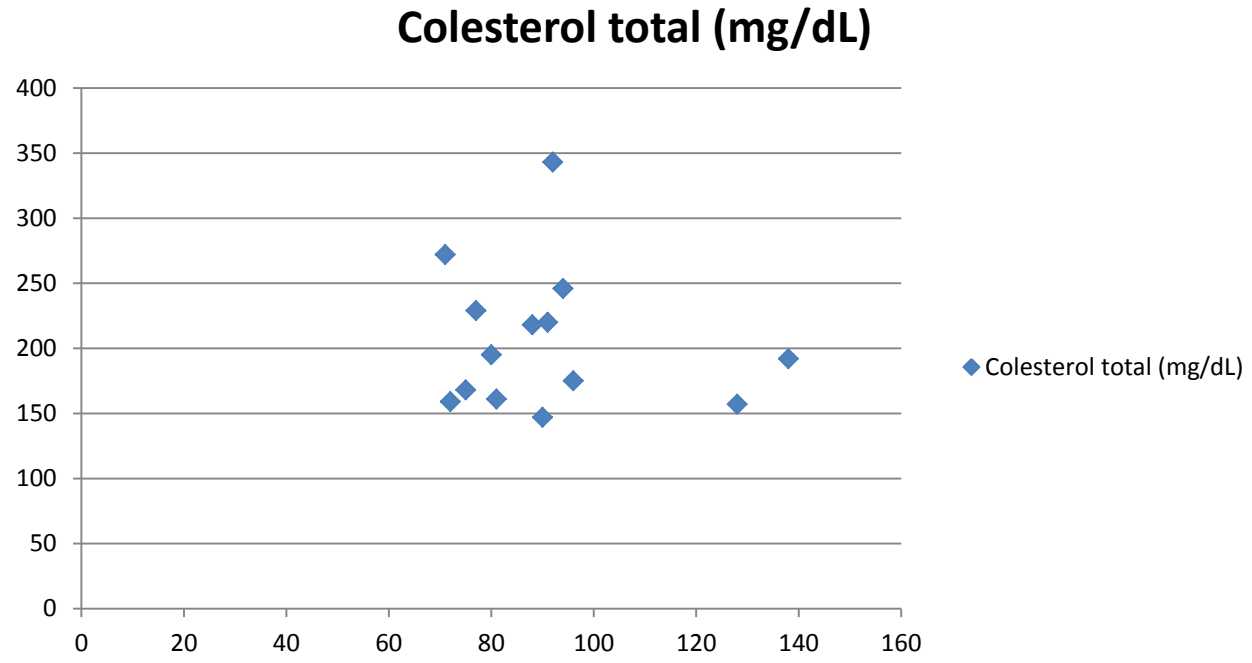
! Vorbim despre relaţie liniară

Graficul de corelaţie – XY Scatter

- Două variabile – pot fi cu două unităţi de măsură diferite
- Variabila dependentă pe axa OY şi variabila independentă pe axa OX

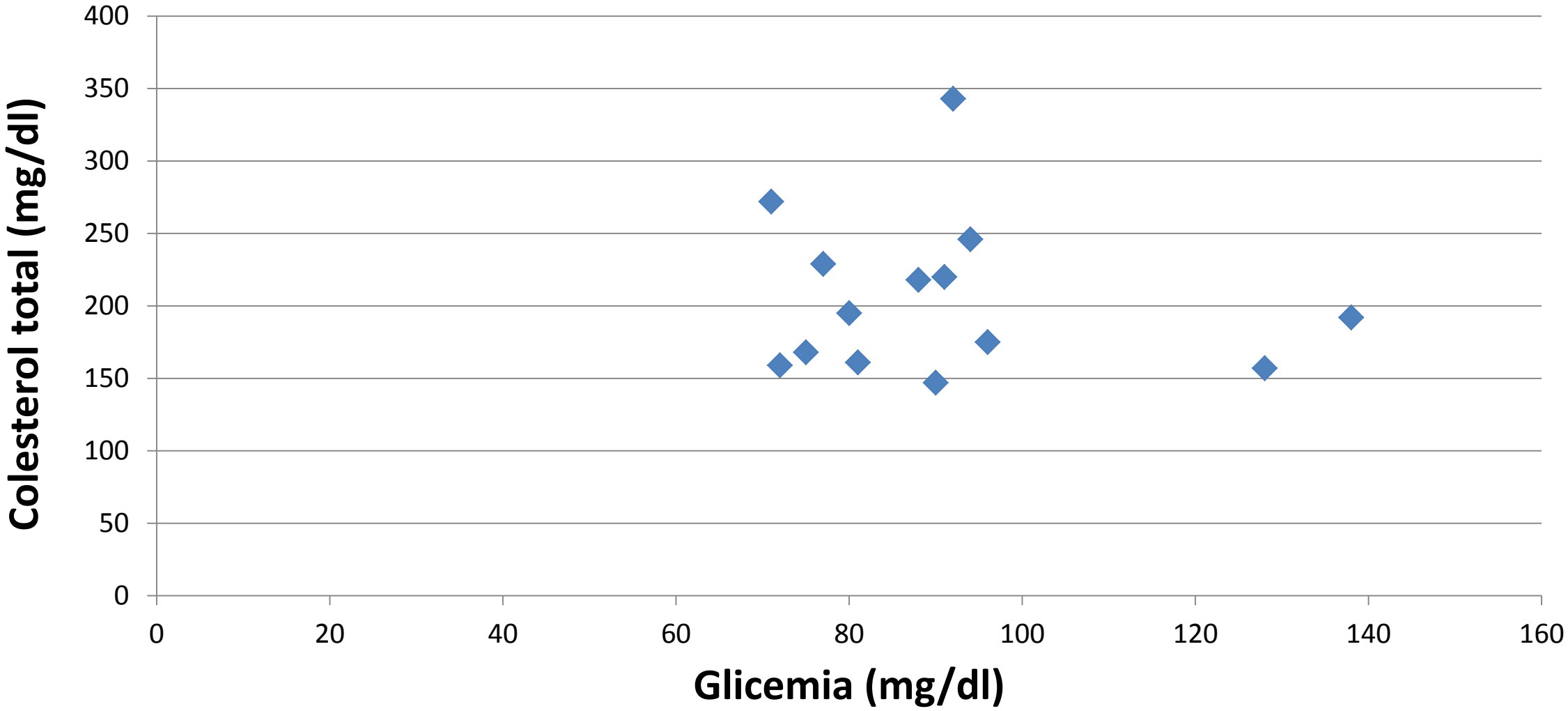
	Glicemie (mg/dL)	Colesterol total (mg/dL)
7	75	168
8	92	343
9	77	229
10	128	157
11	81	161
12	138	192
13	88	218
14	72	159
15	71	272
16	80	195
17	91	220
18	94	246
19	90	147
20	96	175

Selectăm tot tabelul

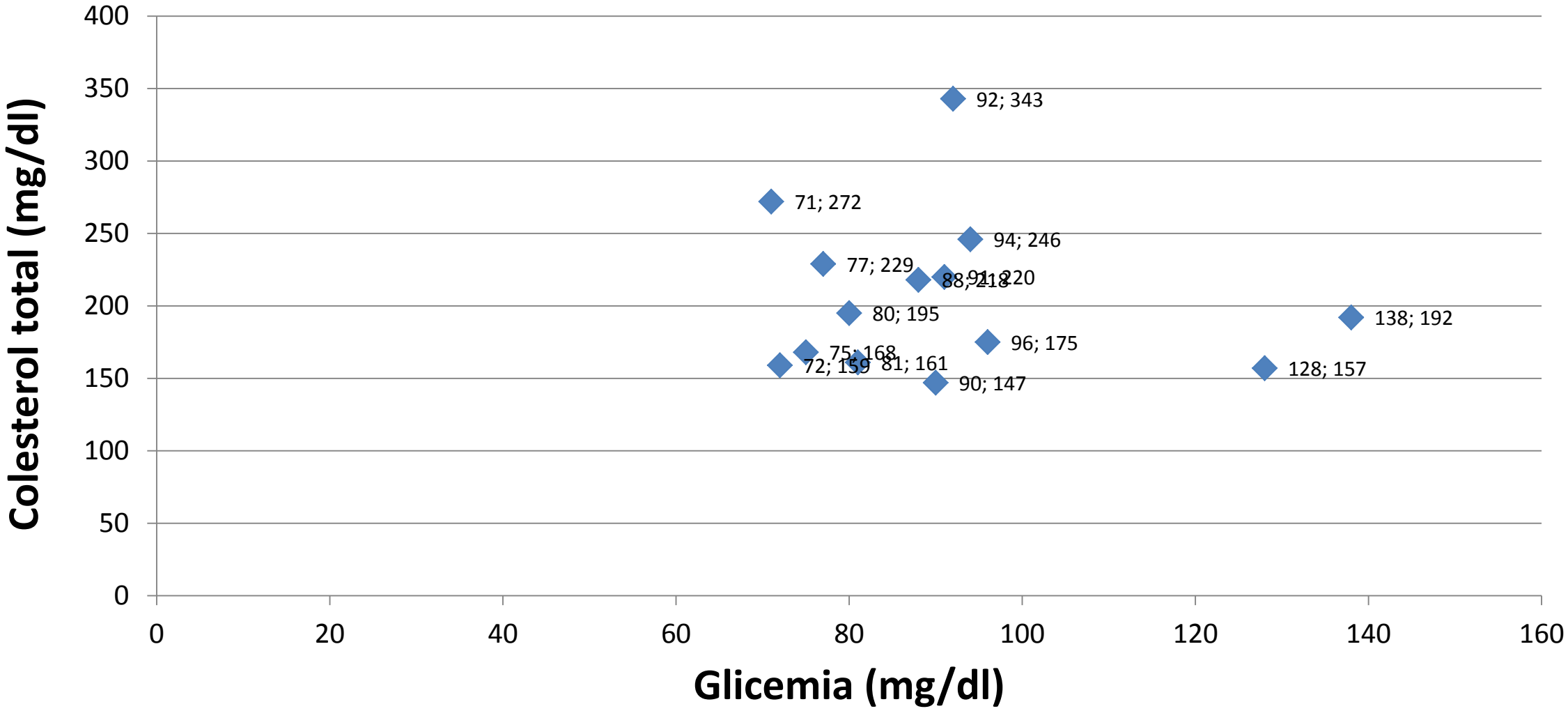


Inserăm grafic Scatter

Corelatia dintre colesterol si glicemie

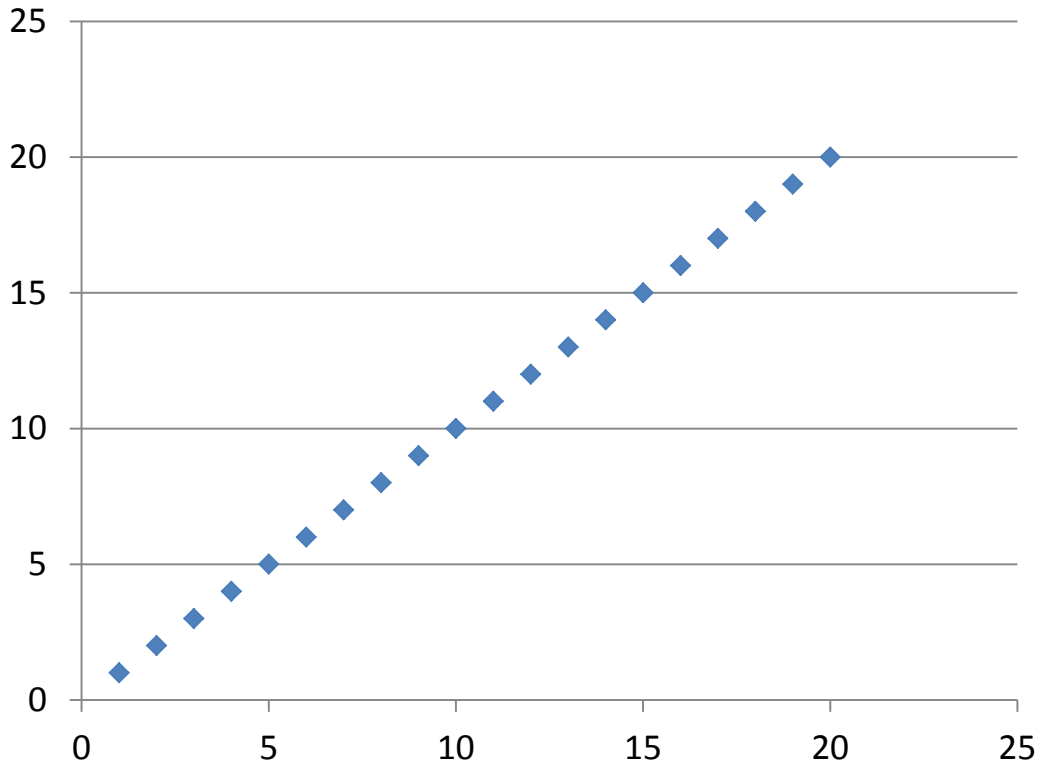


Corelatia dintre colesterol si glicemie

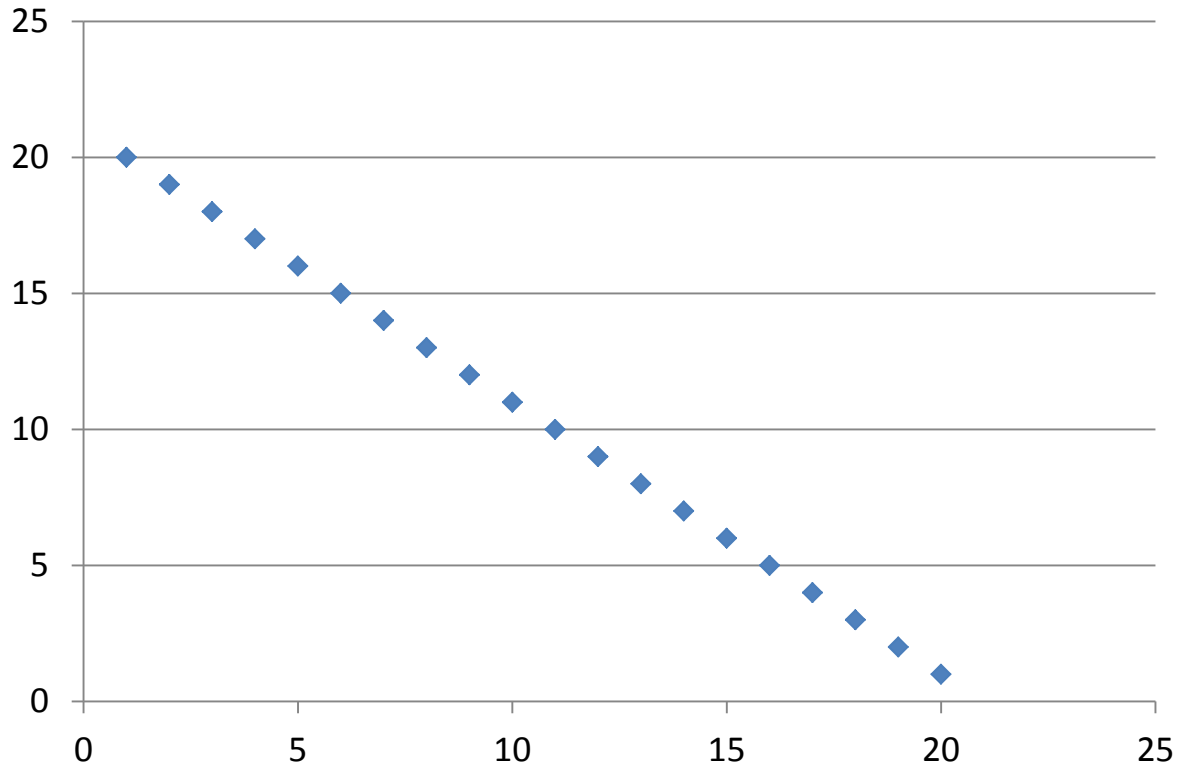


Corelaţia perfectă

- r=1



r=-1



Coeficient de corelaţie Pearson

- Dacă r>0 atunci relaţia dintre X şi Y este **pozitivă** (directă)

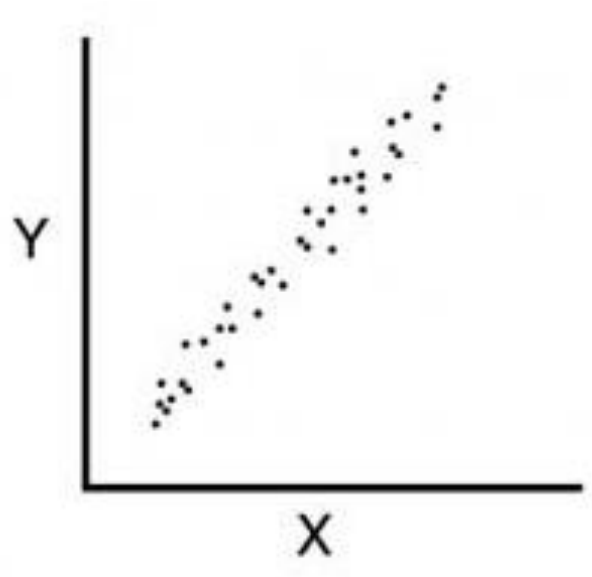
(valorilor mici ale lui X – le corespund valori mici ale lui Y,
valorilor mari ale lui X – le corespund valori mari ale lui Y)

- Dacă r<0 atunci relaţia dintre X şi Y este **negativă** (inversă)

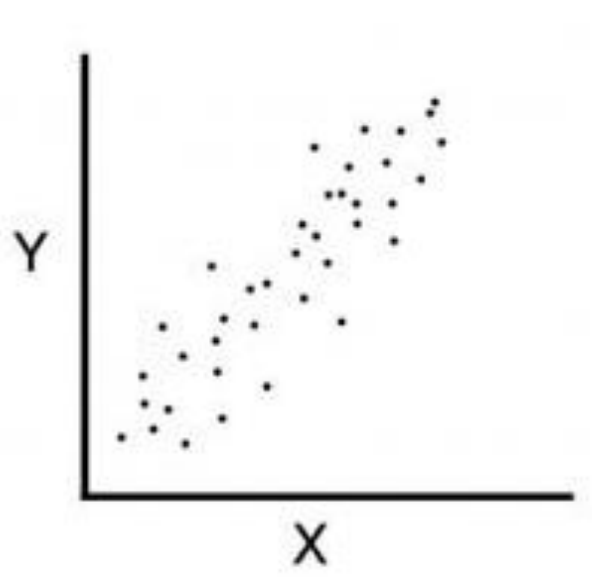
(valorilor mici ale lui X – le corespund valori mari ale lui Y,
valorilor mari ale lui X – le corespund valori mici ale lui Y)

Graficul de corelaţie – XY Scatter

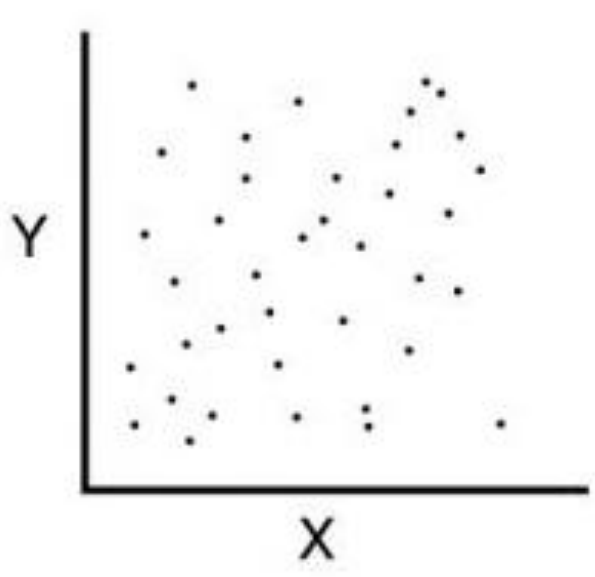
- Relaţie imperfectă, directă, pozitivă



Corelaţie puternică
r≈1



Corelaţie moderată
r≈0.5



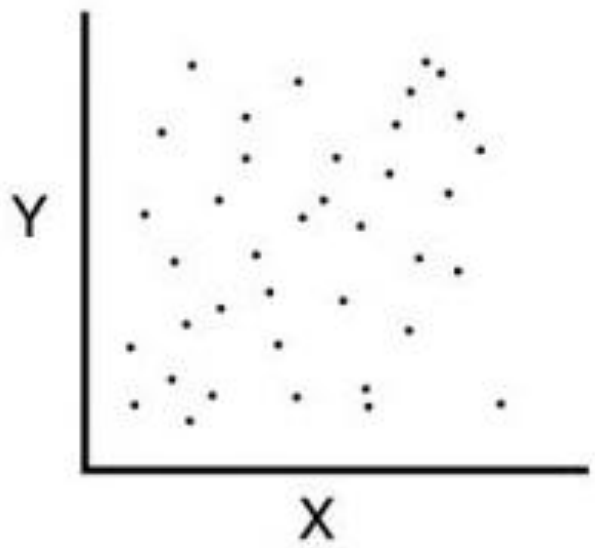
Corelaţie nulă
r≈0

Graficul de corelaţie – XY Scatter

- Relaţie imperfectă, inversă, negativă

Corelaţie puternică
r≈ – 1

Corelaţie moderată
r≈ – 0.5



Corelaţie nulă
r≈0

Regulile lui Colton

- [Colton T. Statistics in Medicine. Little Brown and Company, New York, NY 1974]:
  - $R \subset [-0{,}25\ ;\ +0{,}25]$ → Nu există asociere spre asociere slabă
  - $R \subset (0{,}25\ ;\ +0{,}50] \cup (-0{,}25\ ;\ -0{,}50]$ → Asociere acceptabilă
  - $R \subset (0{,}50\ ;\ +0{,}75] \cup (-0{,}50\ ;\ -0{,}75]$ → Asociere moderată sau bună
  - $R \subset (0{,}75\ ;\ +1] \cup (-0{,}75\ ;\ -1]$ → Asociere foarte bună

Vom reveni la această noţiune în cursurile următoare

Exemplu

- Ex. Corelaţia dintre Greutate şi densitatea minerală osoasă ($r=0.43$, $p<0.001$) la pacienţii cu artroplastie bilaterală totală a genunchiului

- Pozitivă – Greutatea mare – corespunde cu valori mari de densitate osoasă (DMO creşte cu greuatea subiectului)

Regulile lui Colton → relaţie acceptabilă

Ishii Y, Noguchi H, Sato J, Ishii H, Todoroki K, Toyabe SI. Association between body weight and proximal tibial bone mineral density after bilateral total knee arthroplasty. Knee. 2017 Oct;24(5):1153-1159.

Vom reveni la această noţiune în cursurile următoare

Exemplu

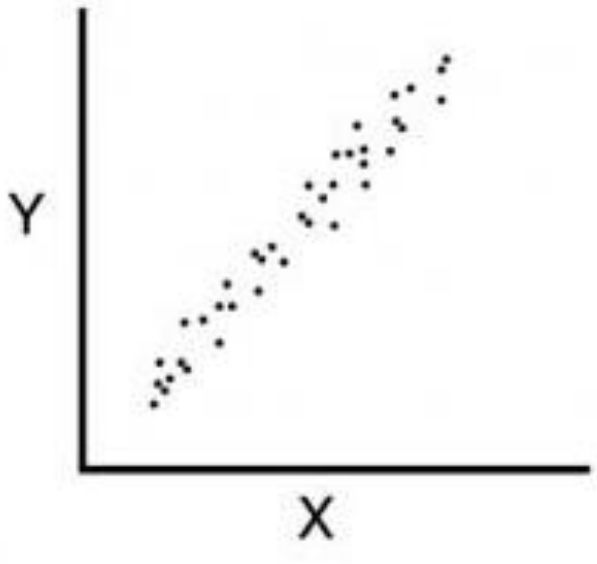
- Ex. Corelaţia dintre Vârsta pacientului şi densitatea minerală osoasă (r= – 0.32, p=0.007) la pacienţii cu artroplastie bilaterala totală a genunchiului

- Pozitivă – Vârstă mare – corespunde cu valori mici de densitate osoasă (DMO scade cu vârsta)

Regulile lui Colton → relaţie acceptabilă

Ishii Y, Noguchi H, Sato J, Ishii H, Todoroki K, Toyabe SI. Association between body weight and proximal tibial bone mineral density after bilateral total knee arthroplasty. Knee. 2017 Oct;24(5):1153-1159.

Graficul de corelaţie – XY Scatter



Interpretare

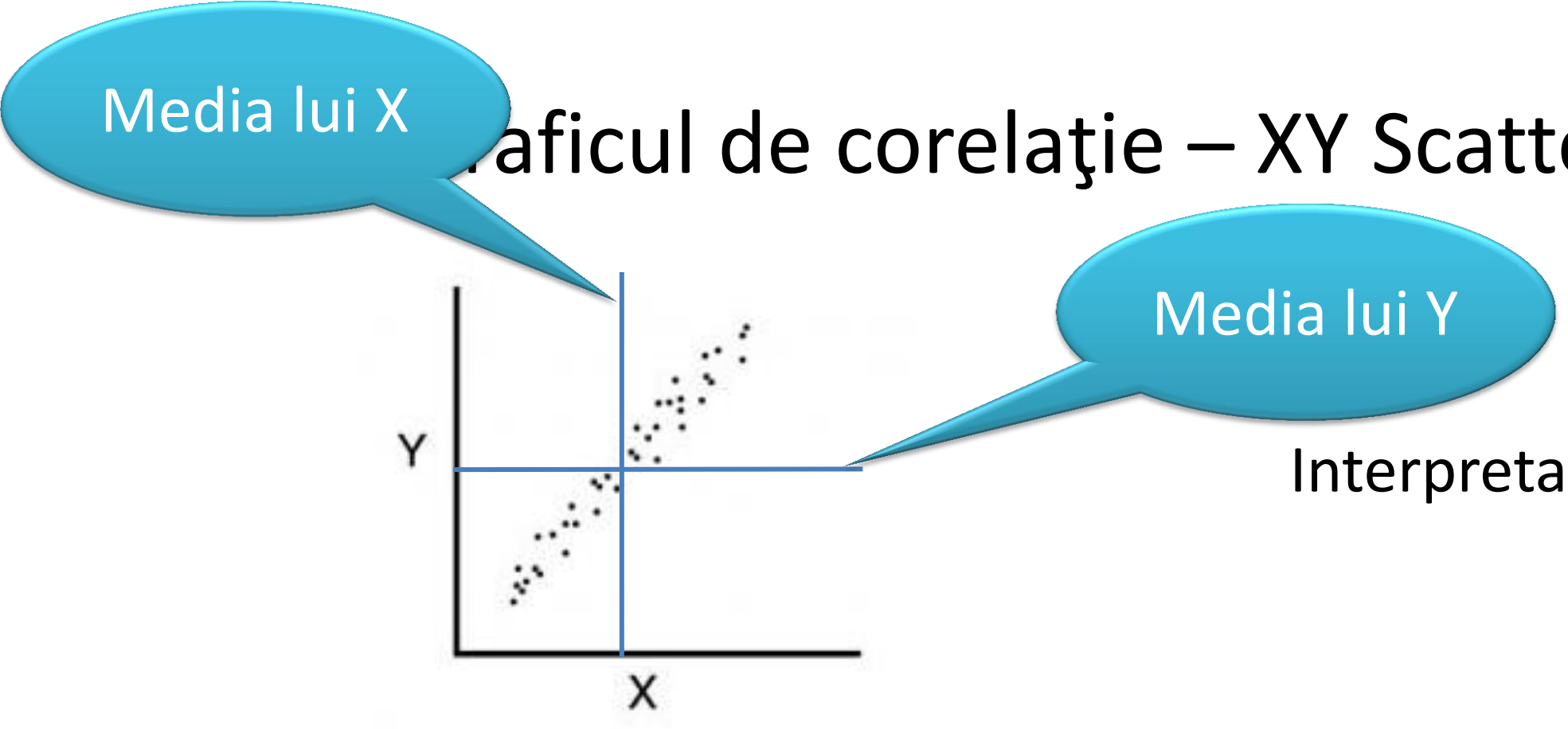
Coeficient de corelaţie r > 0

X creşte, creşte şi Y, dacă r este mare (aproape de 1)

Valori mici ale X corelează cu valori mici ale Y

Valori crescute ale X corelează cu valori crescute ale Y

Graficul de corelaţie – XY Scatter



Interpretare

Coeficient de corelaţie r > 0

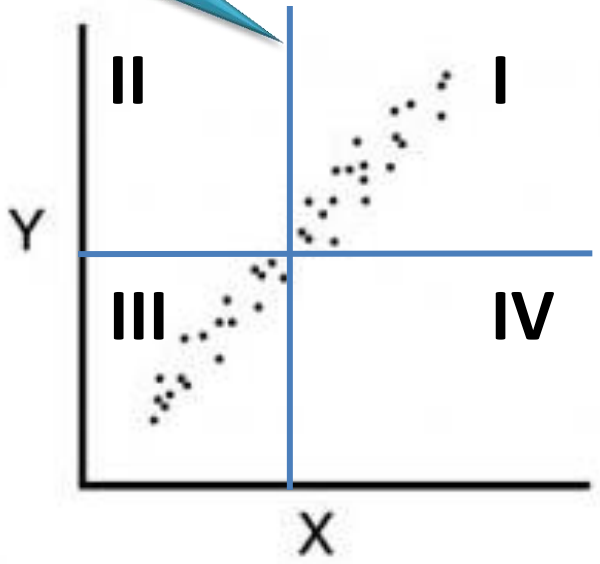
X creşte, creşte şi Y, dacă r este mare (aproape de 1)

Valori mici ale X corelează cu valori mici ale Y

Valori crescute ale X corelează cu valori crescute ale Y

4 cadrane

Graficul de corelaţie – XY Scatter



Interpretare

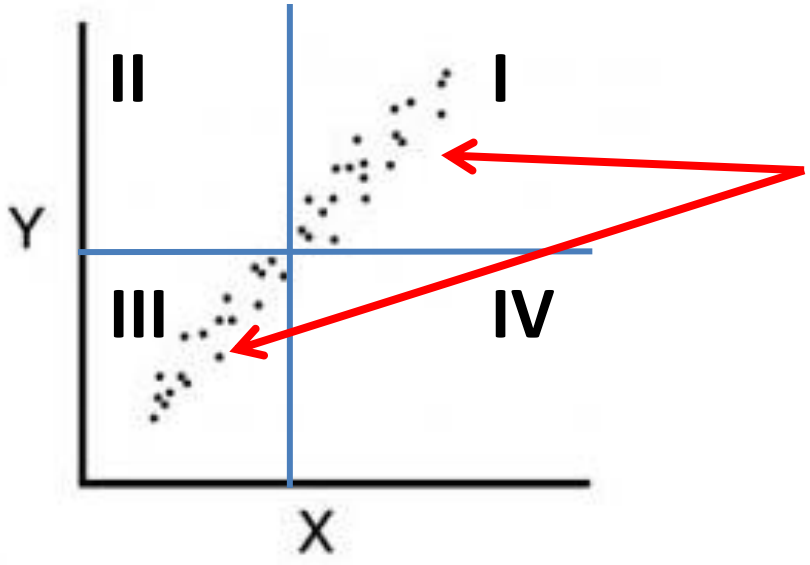
Coeficient de corelaţie r > 0

X creşte, creşte şi Y, dacă r este mare (aproape de 1)

Valori mici ale X corelează cu valori mici ale Y

Valori crescute ale X corelează cu valori crescute ale Y

Graficul de corelaţie – XY Scatter



Când punctele sunt în cadranul I şi III atunci corelaţia este puternică

!

Coeficient de corelaţie r > 0

X creşte, creşte şi Y, dacă r este mare (aproape de 1)

Valori mici ale X corelează cu valori mici ale Y

Valori crescute ale X corelează cu valori crescute ale Y

Graficul de corelaţie – XY Scatter

Interpretare

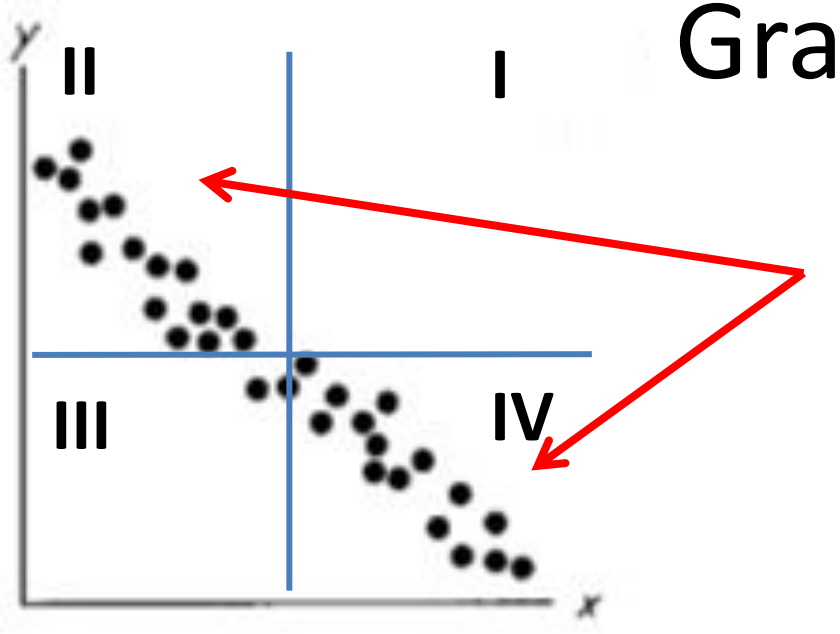
Coeficient de corelaţie $r < 0$

X creşte şi Y descreşte, dacă r este mic (aproape de -1)

Valori mici ale X corelează cu valori crescute ale Y

Valori crescute ale X corelează cu valori mici ale Y

Graficul de corelaţie – XY Scatter



Când punctele sunt în cadranul II şi IV atunci corelaţia este puternică

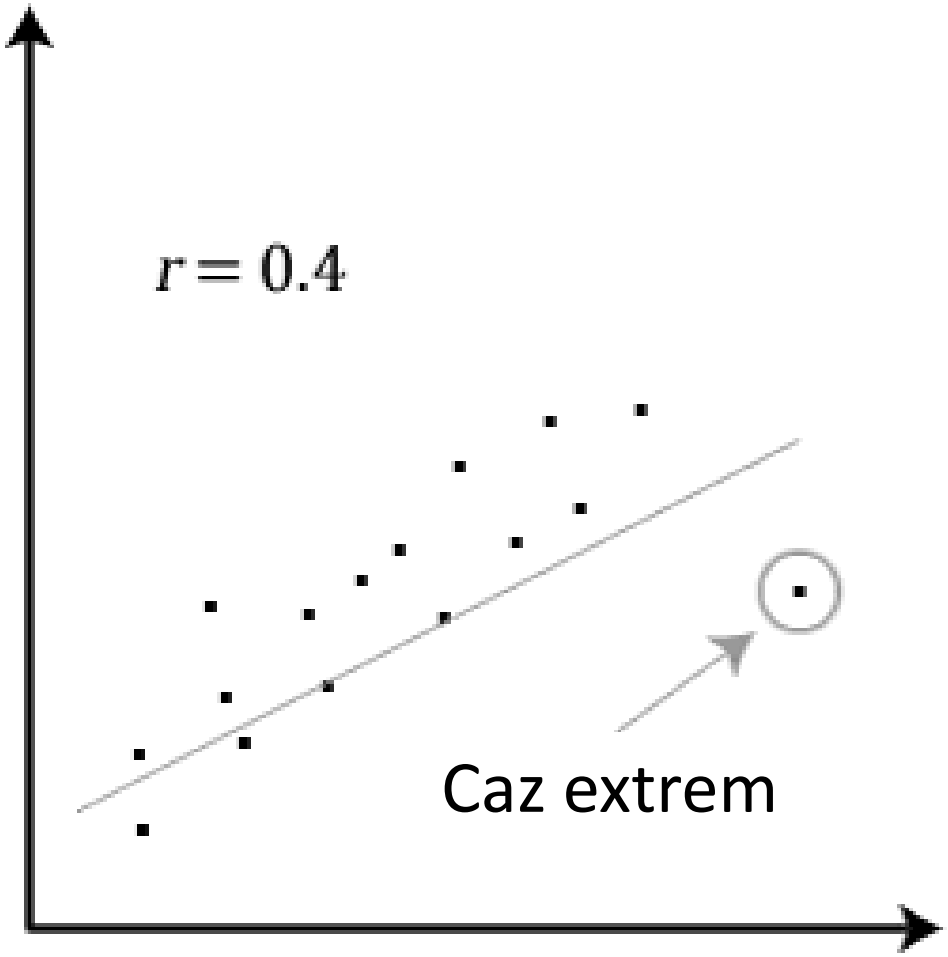
Coeficient de corelaţie r < 0

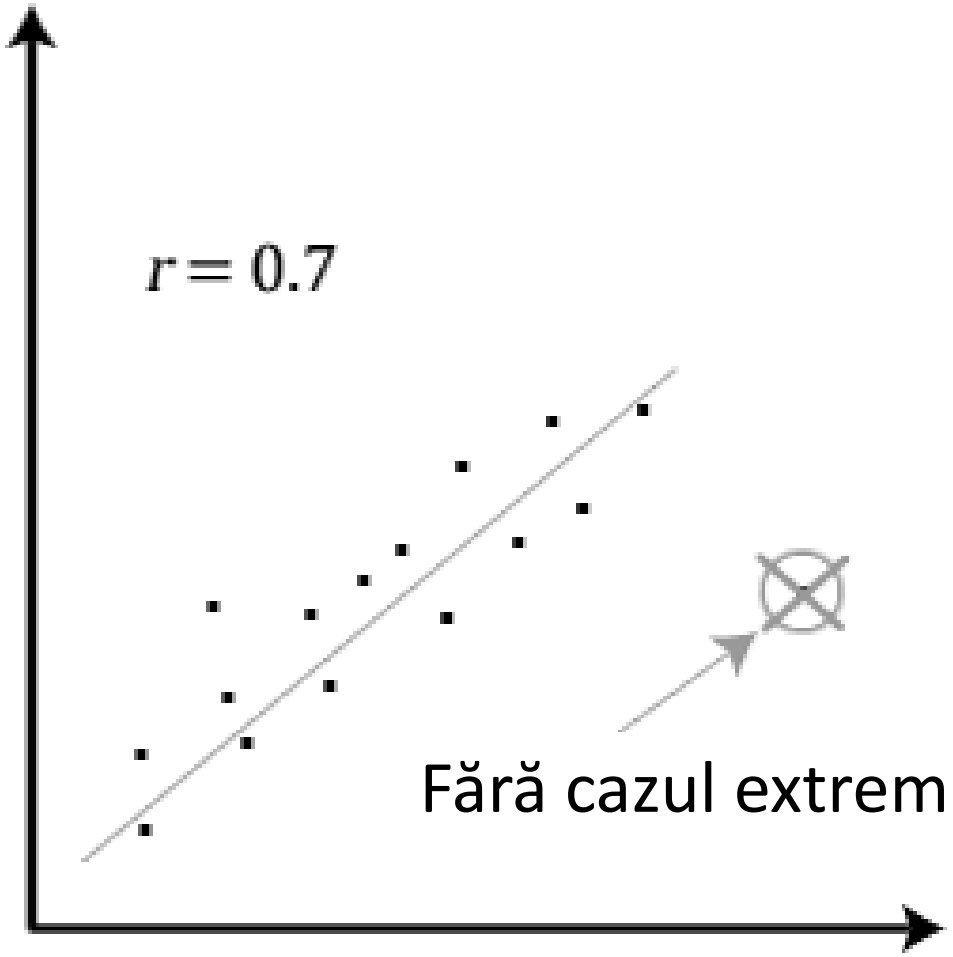
X creşte şi Y descreşte, dacă r este mic (aproape de -1)

Valori mici ale X corelează cu valori crescute ale Y

Valori crescute ale X corelează cu valori mici ale Y

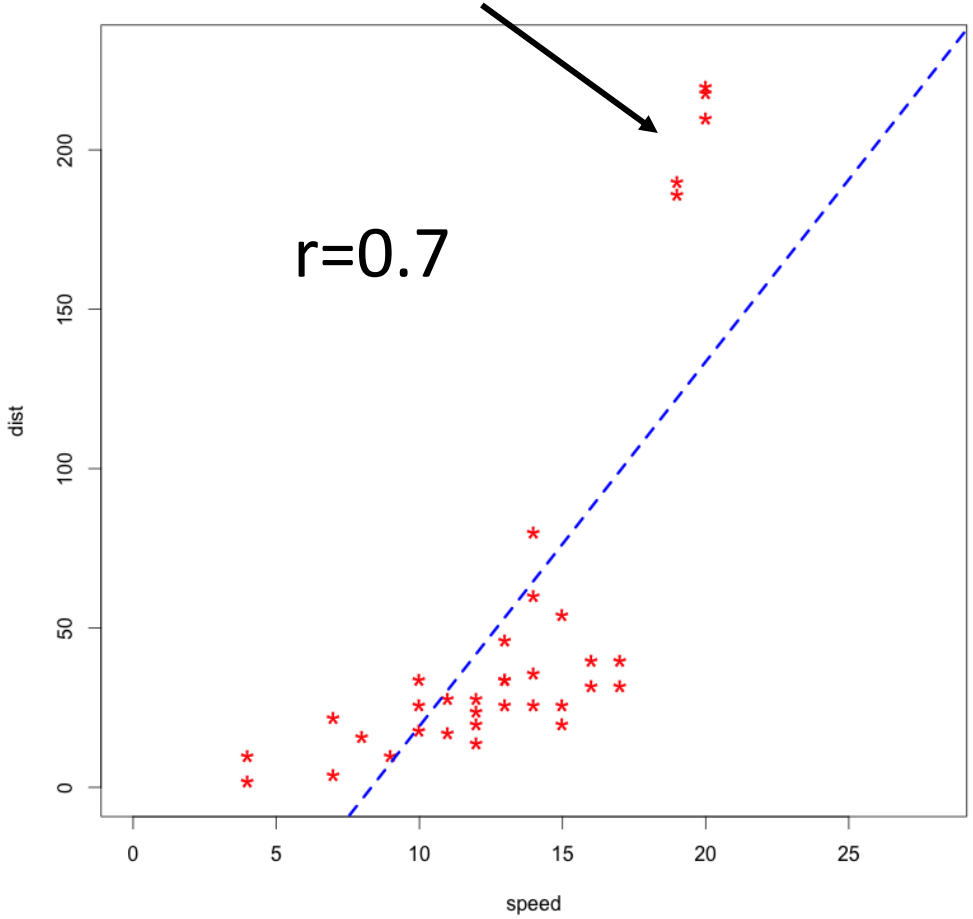
Efectul cazurilor extreme asupra r – coeficientul de corelație Pearson



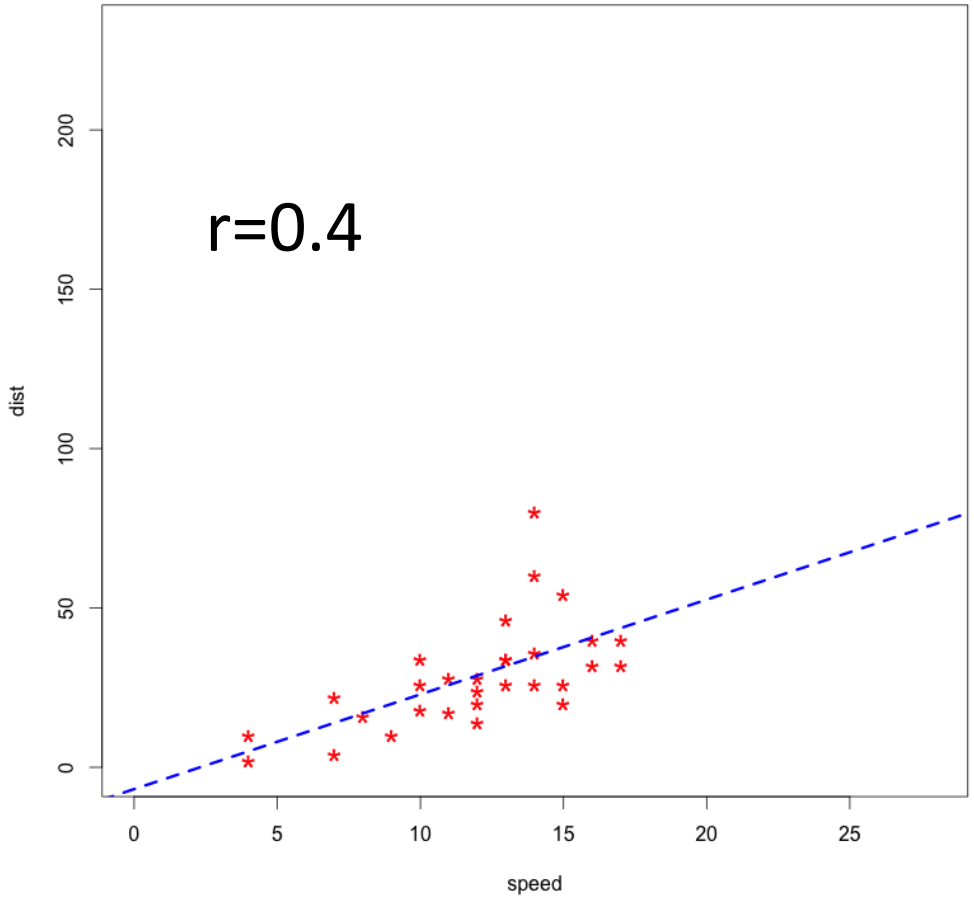


Efectul cazurilor extreme asupra r – coeficientul de corelație Pearson

Cazuri extreme



Fără cazurile extreme



Mărimea efectului – coeficientul de determinare d

$$d=r^2$$

Ex. r=0.5, atunci d=0.25

25% din variația lui Y poate fi explicat prin relația liniară cu X

Cerințe pentru calcularea coeficientului de corelaţie Pearson

- X şi Y să nu aibă cazuri extreme
- X şi Y să fie variabile cantitative
- Relaţia să fie liniară

- Soluţii posibile:
  - Transformări: logaritmare, normalizare
  - Scoaterea cazurilor extreme din analiză
  - calcularea coeficientului de corelaţie Spearman

Graficul de corelaţie – XY Scatter

Relaţie neliniară

- Nu există o relație liniară între X și Y – asocierea respectivă nu este adecvată pentru calcularea coeficientului de corelație Pearson, ar trebui să calculăm coeficientul de corelație Spearman

Coeficientul de corelaţie Spearman

- Fie R_X şi R_Y rangurile a două variabile X şi Y:

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(R_{X,i} - \overline{R_X}\right)\left(R_{Y,i} - \overline{R_Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_{X,i} - \overline{R_X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(R_{Y,i} - \overline{R_Y}\right)^2}}$$

- Poate fi calculat şi pentru două variabile ordinale

Coeficientul de corelaţie Spearman

- Indică asocierea dintre X şi Y
- Întotdeauna între -1 şi 1

$$r \in [-1,1]$$

când $|\ r\ |$ se apropie de 1 asocierea este mai mare

când $|\ r\ |$ se apropie de 0 asocierea este mai mică

! Vorbim despre relaţie neliniară

Coeficientul de corelaţie Spearman

- Dacă r>0 atunci asocierea dintre X şi Y este **pozitivă** (directă, neliniară)
- Dacă r<0 atunci asocierea dintre X şi Y este **negativă** (inversă, neliniară)
- Se interpretează după regulile lui Colton

Regresie

- Predicţie = regresie – modelarea matematică a relației dintre Y și X
- univariată - modelul cu o singură variabilă independentă X
- multivariată - modelul cu mai multe variabilă independentă X_i, i=2,n

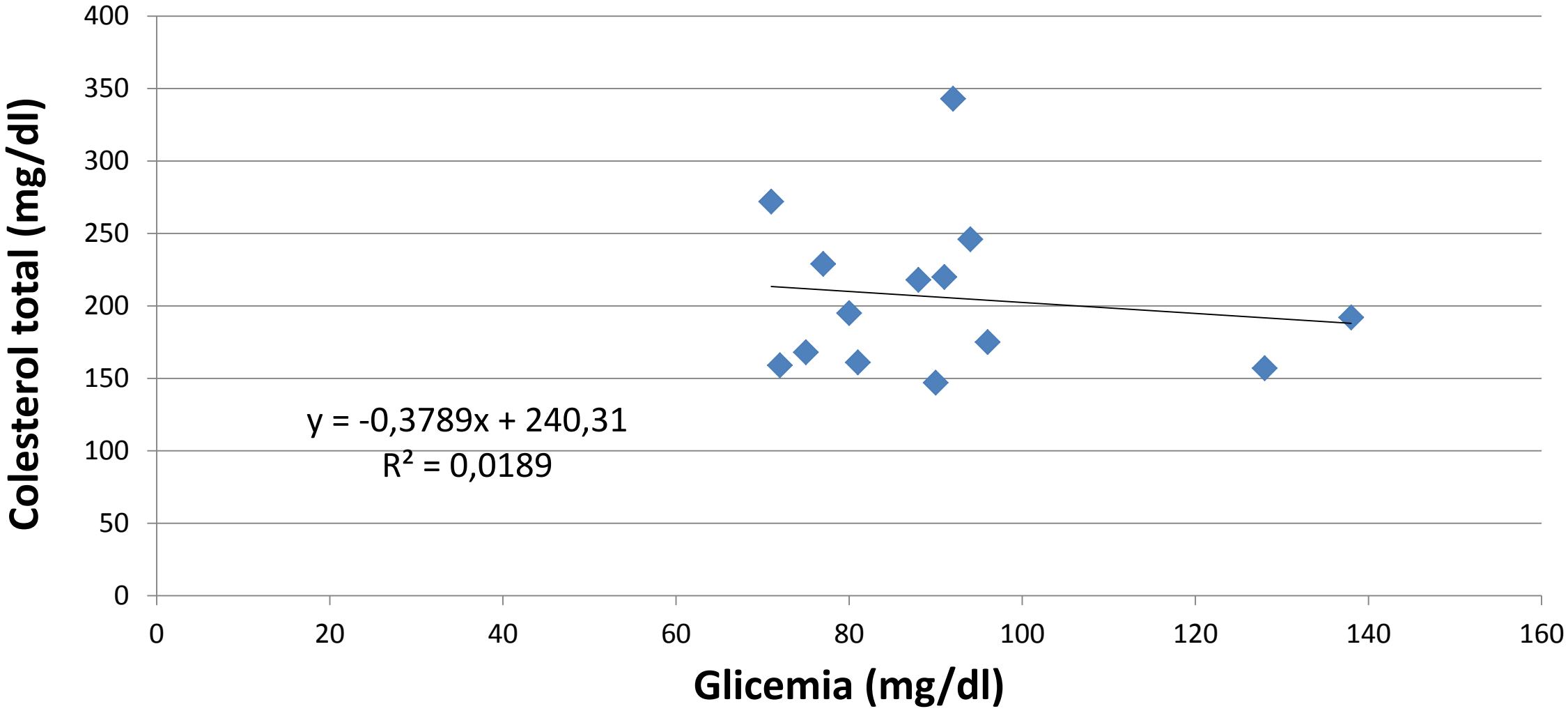
Regresie liniară

- Regresie liniară - funcția de predicție liniară: $Y=aX+b$

- **Metoda celor mai mici pătrate** - pentru a obține ecuația matematică - modalitatea de a determina ecuația celei mai potrivite linii pentru norul de date
- Punctul în care linia traversează axa Y este notat cu a și panta liniei cu b, ecuația dreptei:

$$Y=aX+b$$

Corelatia dintre colesterol si glicemie



Regresie liniară (Y=aX+b)

- Y creşte cu **a** de câte ori X creşte cu 1

$$Y = 3X + 5$$

- X=1, atunci $Y = 3 \cdot 1 + 5 = 8$
- X=2, atunci $Y = 3 \cdot 2 + 5 = 11$
- X=3, atunci $Y = 3 \cdot 3 + 5 = 14$
- ….
- Dacă **a** este pozitiv, atunci când X crește, Y crește

Regresie liniară (Y=aX+b)

- Y creşte cu -**a** de câte ori X creşte cu 1

$$Y = -3X + 5$$

- X=1, atunci $Y = -3 \cdot 1 + 5 = 2$
- X=2, atunci $Y = -3 \cdot 2 + 5 = -1$
- X=3, atunci $Y = -3 \cdot 3 + 5 = -4$
- ….
- Dacă **a** este negativ, când X creşte Y descreşte

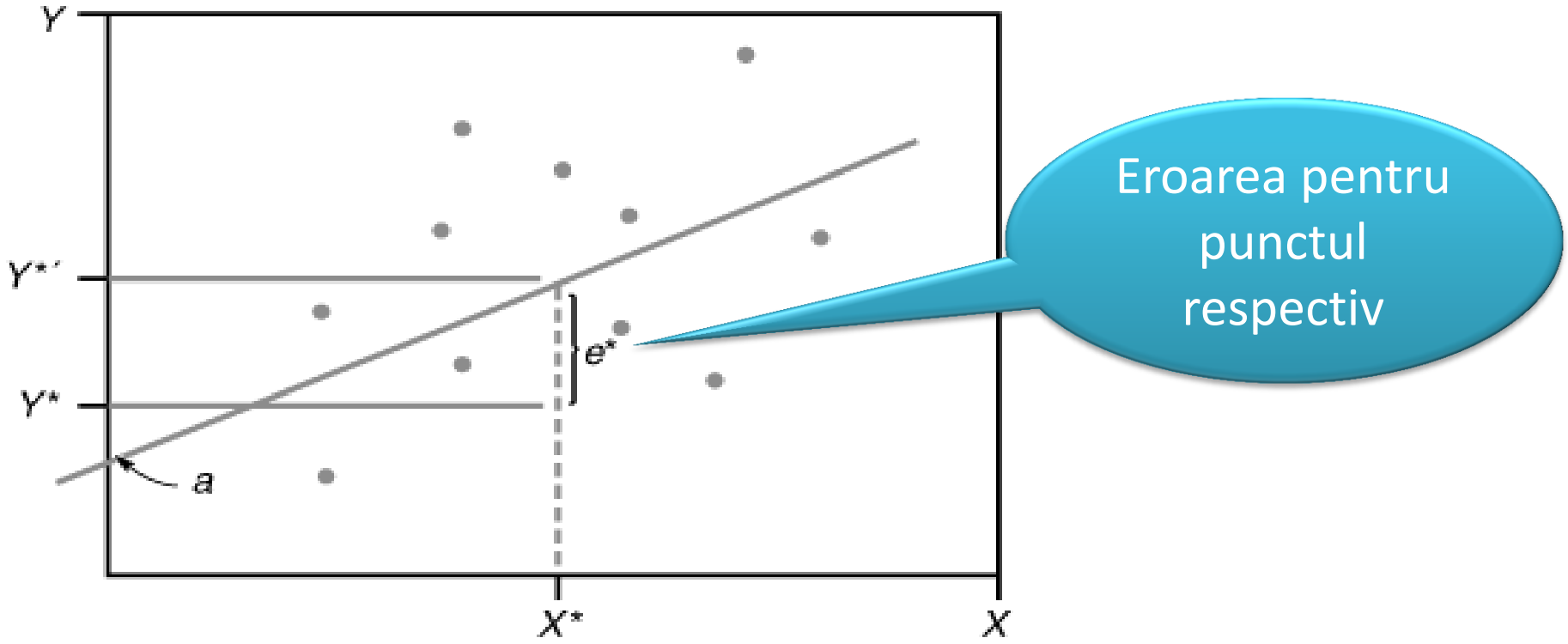
- Formula pentru coeficienții liniei de regresie:

$$b = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(X - \overline{X})^2}$$

$$a = \overline{Y} - b\overline{X}$$

Regresie liniară (Y=aX+b)

- Când modelul liniar nu se potrivește perfect (nu toate punctele sunt pe linia de regresie) --> avem o eroare



Cum realizăm predicţii?

- Avem nevoie de ecuaţia de regresie

Cum putem obţine ecuaţia de regresie?

- Ex. Greutatea creşte cu vârsta

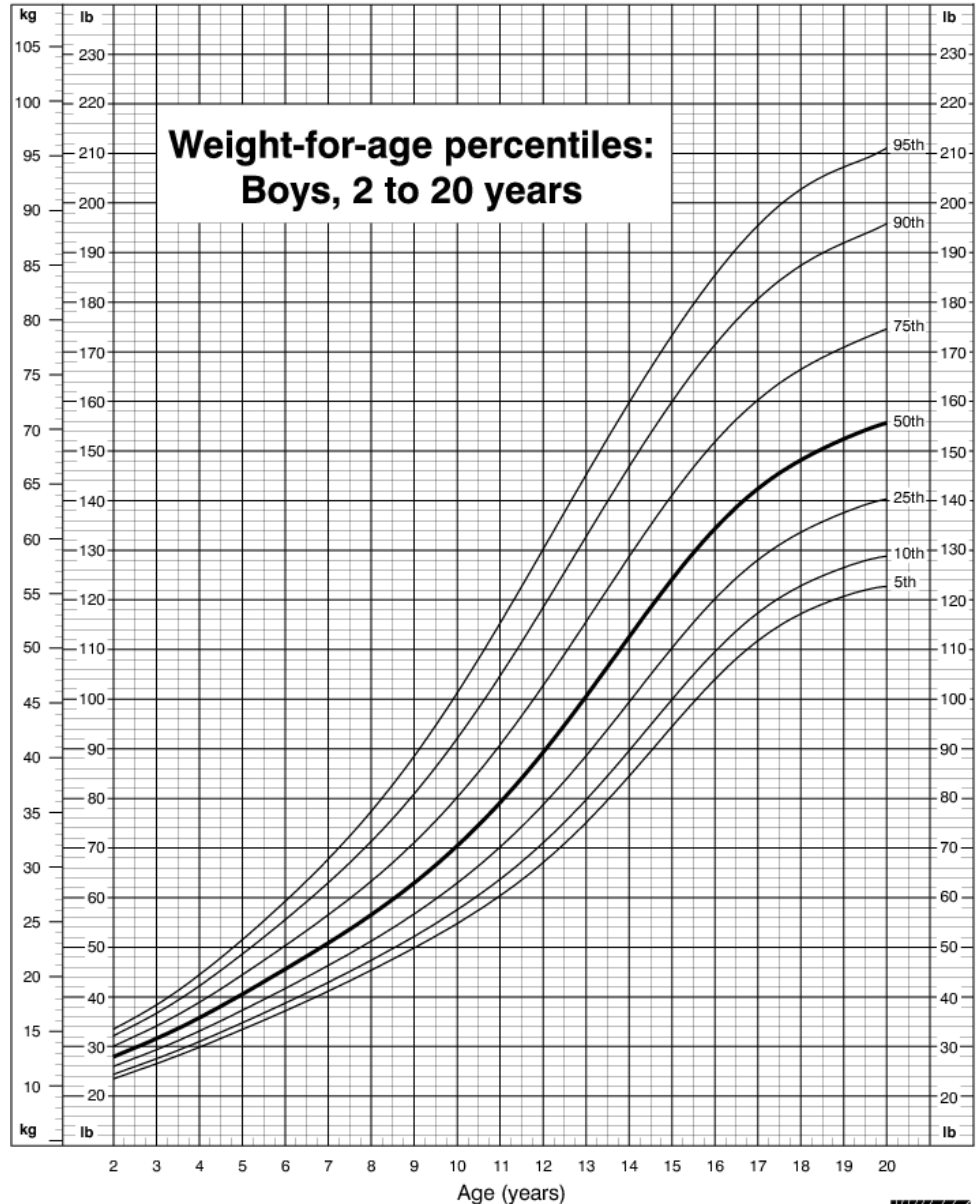
- Greutatea variază de la 1 la 200
- Vârsta variază de la 0 la 110

- Considerăm vârsta. Pentru fiecare valoare a vârstei luăm minim 3 indivizi. Ca să ne asigurăm că avem minim 3 indivizi pe fiecare valoare a vârstei selectăm la întâmplare în loc de 330 de indivizi: 3000. Obţinem greutatea fiecărui individ.

Ce constatăm?

- Femeile şi bărbaţii au greutăţi diferite
- Copiii şi adolescenţii au dependenţe diferite faţă de adulţi
- Este nevoie să renunţăm la unii indivizi
- Relaţia nu este liniară
- Ne focusăm numai pe adolescenţi şi copii, băieţi

CDC Growth Charts: United States



Published May 30, 2000.
SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).

SAFER • HEALTHIER • PEOPLE™

Acum să realizăm predicţii

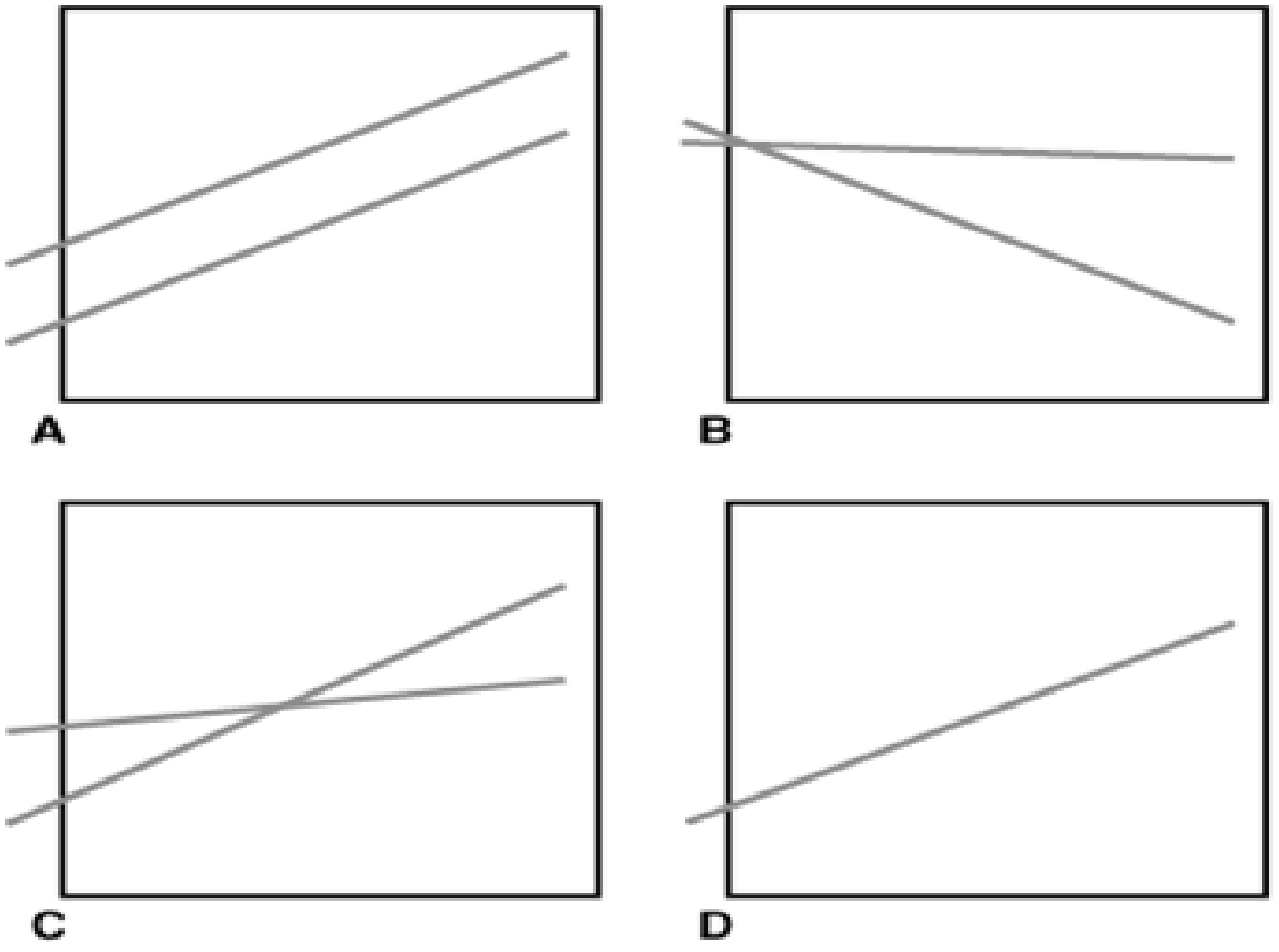
- Câte kg va avea un băiat la 10 ani?
  - Între 29 kg şi 36 kg cu 50% eroare
  - Între 24 kg şi 46 kg cu 5% eroare
- Câte kg va avea un băiat la 5 ani?
  - Între 17 kg şi 19 kg cu 50% eroare
  - Între 15 kg şi 24 kg cu 5% eroare

Regresie liniară

- Eroarea standard – media erorilor de predicţie pentru fiecare punct în parte

$$ES_{X,Y}=\sqrt{\frac{\sum_{i=1}^{n}(Y_{i-predicted}-Y_{i-observed})^2}{n-2}}$$

Dacă repetăm studiul, linia de regresie este diferită



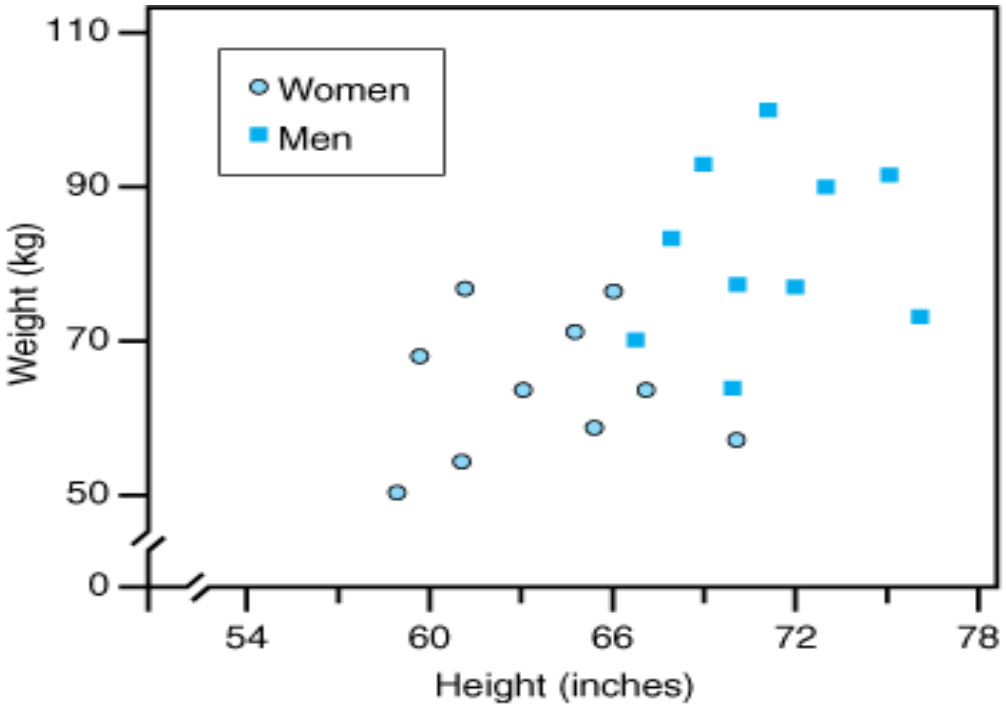
- A. Pantă egală, constantă diferită; B. Pantă diferită, constantă egală;
- C. Pantă diferită și constantă diferită; D. Aceeași constantă și pantă

Relaţie neliniară - ce putem face?

- Transformarea datelor
- Regresie polinomială
- Regresie pe mai multe secțiuni (intervale)

Erori în utilizarea regresiei

- Amestecarea a două populații diferite
- Amestecarea datelor rezultate din măsurători repetate pe aceeași axă (Y nu are toate datele independente)



Coeficientul de corelaţie sau regresia liniară ?

- Coeficientul de corelaţie este independent de unitatea de măsură, regresia liniară nu
- Atunci când greutatea este măsurată în kg sau grame, linia de regresie are un coeficient diferit, în timp ce coeficientul de corelație este același

- Coeficientul de corelație și coeficientul liniei de regresie au același semn (+ sau -)

- Nu are sens calcularea coeficientul de corelație între X și Y dacă Y este calculat din X
- Ex. IMC este corelat cu înălțimea și greutatea (a fost calculat pe baza lor - avem deja ecuaţia nu mai este necesar să o calculăm)
- $X \cdot Y \cdot Z$ se corelează cu Z
- X / Z se corelează cu Y / Z

Corelaţia nu este cauzalitate

- În Cluj-Napoca s-a observat o corelare între creşterea numărului de fracturi ale picioarelor şi creşterea numărului de filme vizionate

- Primarul urmează să interzică firmele furnizoare de televiziune prin cablu şi internet

Corelaţia nu este cauzalitate

- De fapt există o altă variabilă care influnţează cele două creşteri concomitente:
- frigul care se instalează odată cu venirea toamnei şi a iernii,

acesta duce

  - la apariţia gheţii care înmulţeşte numărul de fracturi
  - oamenii stau mai mult în casă şi se uită la mai multe filme

Corelaţia nu este cauzalitate

- De la Ecuator la Pol scade temperatura, iar această scădere se corelează cu creşterea înălţimii persoanelor care locuiesc în zonele respective, astfel că s-a ajuns la concluzia că înălţimea persoanelor depinde de temperatura mediului

Corelaţia nu este cauzalitate

- La Ecuator locuiesc pigmeii - un trib de oameni de înălţime mică
- Europa este locuită de europeni - o populaţie de înălţime medie
- În nordul Europei trăiesc persoane mai înalte (urmaşii vikingilor)

- O coincidenţă: înălţimea persoanelor depinde de temperatura mediului

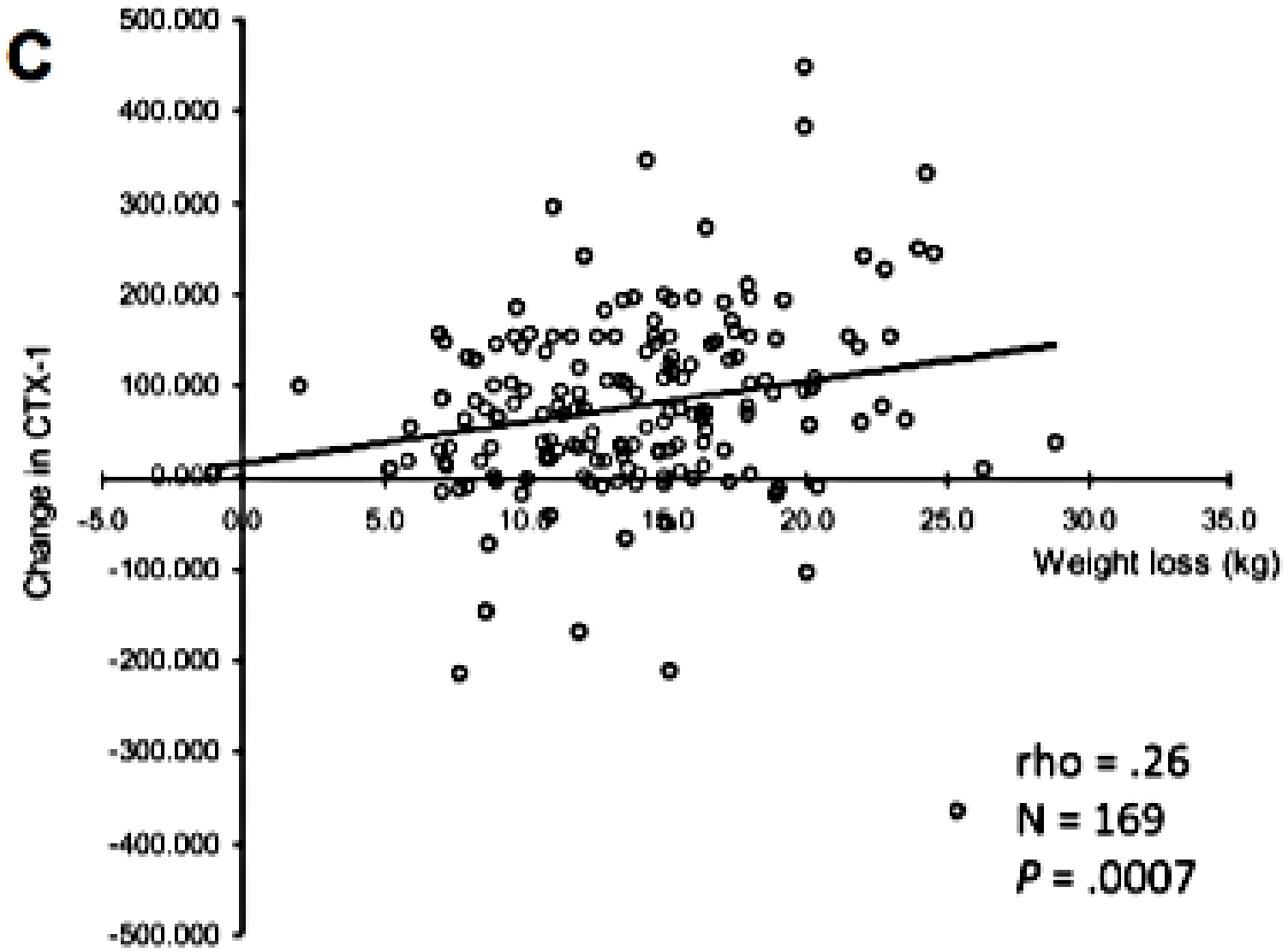
- Corelaţia trebuie să aibă şi sens.

Scenariu

- 192 obezi cu osteoartrita genunchiului (OA)
- OA evaluat cu un biomarker uCTX-I (Urine C-terminal telopeptide of collagen type I)
- Program intensiv de scădere în greutate

- Scopul studiului: scăderea în greutate se asociază cu creşterea biomarkerului uCTX-I
- uCTX-I a fost direct corelată cu scăderea în greutate ($r = 0.22$, $P = 0.0007$)

Bartels EM, Christensen R, Christensen P, Henriksen M, Bennett A, Gudbergsen H, Boesen M, Bliddal H. Effect of a 16 weeks weight loss program on osteoarthritis biomarkers in obese patients with knee osteoarthritis: a prospective cohort study. Osteoarthritis Cartilage. 2014 Nov;22(11):1817-25.



Bartels EM, Christensen R, Christensen P, Henriksen M, Bennett A, Gudbergsen H, Boesen M, Bliddal H. Effect of a 16 weeks weight loss program on osteoarthritis biomarkers in obese patients with knee osteoarthritis: a prospective cohort study. Osteoarthritis Cartilage. 2014 Nov;22(11):1817-25.

Scenariu

- Scop - asocierea dintre durerea resimţită la nivelul genunchiului și semnele de inflamație în padul de grăsime infrapatelar (IPFP) la pacienții obezi cu osteoartrita genunchiului

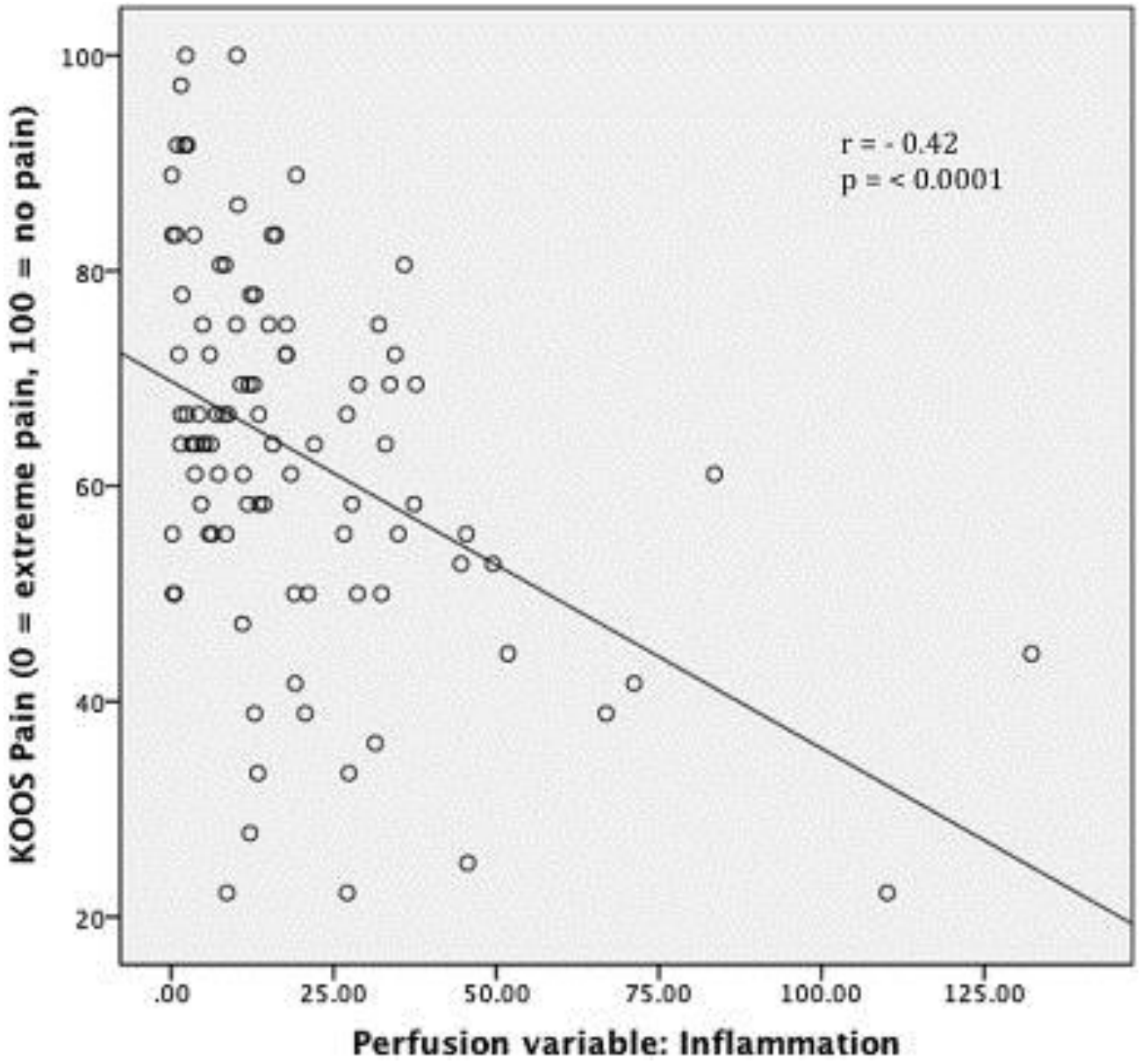
- Inflamaţia în padul de grăsime infrapatelar evaluată cu RMN (CE-MRI şi DCE-MRI)
- KOOS - scorul final al osteoartritei (durerea și alte simptome) 100 = fără durere, 0 = durere extremă
- 95 de pacienţi

[Ballegaard C, Riis RG, Bliddal H, Christensen R, Henriksen M, Bartels EM, Lohmander LS, Hunter DJ, Bouert R, Boesen M. Knee pain and inflammation in the infrapatellar fat pad estimated by conventional and dynamic contrast-enhanced magnetic resonance imaging in obese patients with osteoarthritis: a cross-sectional study. Osteoarthritis Cartilage. 2014 Jul;22(7):933-40.]

- Coeficient de corelaţie

Spearman

- r=-0,42



[Ballegaard C, Riis RG, Bliddal H, Christensen R, Henriksen M, Bartels EM, Lohmander LS, Hunter DJ, Bouert R, Boesen M. Knee pain and inflammation in the infrapatellar fat pad estimated by conventional and dynamic contrast-enhanced magnetic resonance imaging in obese patients with osteoarthritis: a cross-sectional study. Osteoarthritis Cartilage. 2014 Jul;22(7):933-40.]

- Mulţumesc!