Corelații și regresii

Relații între două sau mai multe variabile observate pe același eșantion

Obiective

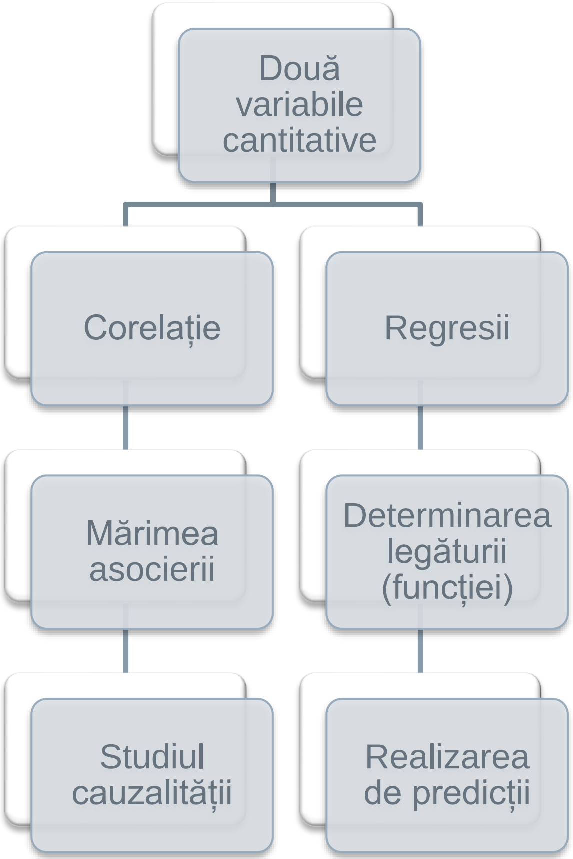
- Evaluarea grafică a relației dintre două variabile cantitative
- Indici de corelație
- Regulile lui Colton
- Teste statistice pentru verificarea semnificației
- Regresia

STATISTICI DESCRIPTIVE ÎN DOUĂ DIMENSIUNI

- Vârsta: X: $X_1, X_2, \ldots, X_n$
- TAS: Y: $Y_1, Y_2, \ldots, Y_n$.

- Să se stabilească dacă **există o legătură** între variabilele X şi Y şi să se determine o modalitate de a **măsura intensitatea** acestei legături.

STATISTICI DESCRIPTIVE ÎN DOUĂ DIMENSIUNI

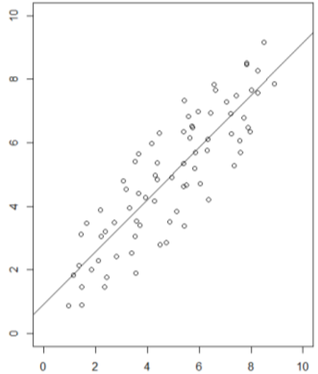


Studiul relației (legăturii) dintre 2 variabile cantitative

- 0. Graficul scatter
- 1. Intensitatea și direcția relației
  - Coeficientul de corelație (Pearson sau Spearman)
  - Coeficientul de determinare
  - Panta regresiei (coeficientul b_i al variabilelor independente)
- 2. Predicție: prezice valorile unei variabile cunoscând valorile celeilalte
  - Regresia - determina o funcție (funcție de regresie) astfel încât Y = f (X)
- 3. Generalizarea la nivelul populației – test statistic

Graficului scatter (nor de puncte)
Evaluarea existenței unei relații liniare

- ▷ Distribuția punctelor pe grafic pare să sugereze că există o tendință ca ele să se afle în jurul unei linii imaginare (pot fi aproximate printr-o dreaptă) - relația poate fi liniară



- ▷ Distribuția punctelor pe grafic NU pare să sugereze că există o tendință ca ele să se afle în jurul unei linii imaginare (NU pot fi aproximate printr-o dreaptă) - relația nu este liniară (poate fi relație exponențială, logaritmică etc)

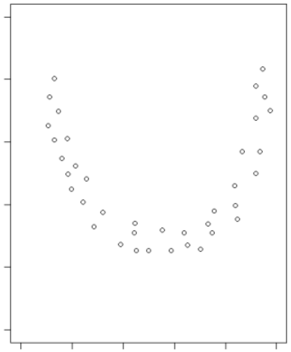


DIAGRAMA DE DISPERSIE

- Diagrama de dispersie asociată unei tabel de date bidimensional:
  - X: X1, X2,..., Xn
  - Y: Y1, Y2,..., Yn

- se obţine reprezentând grafic punctele de coordonate (Xi,Yi) i=1,2,...,n.

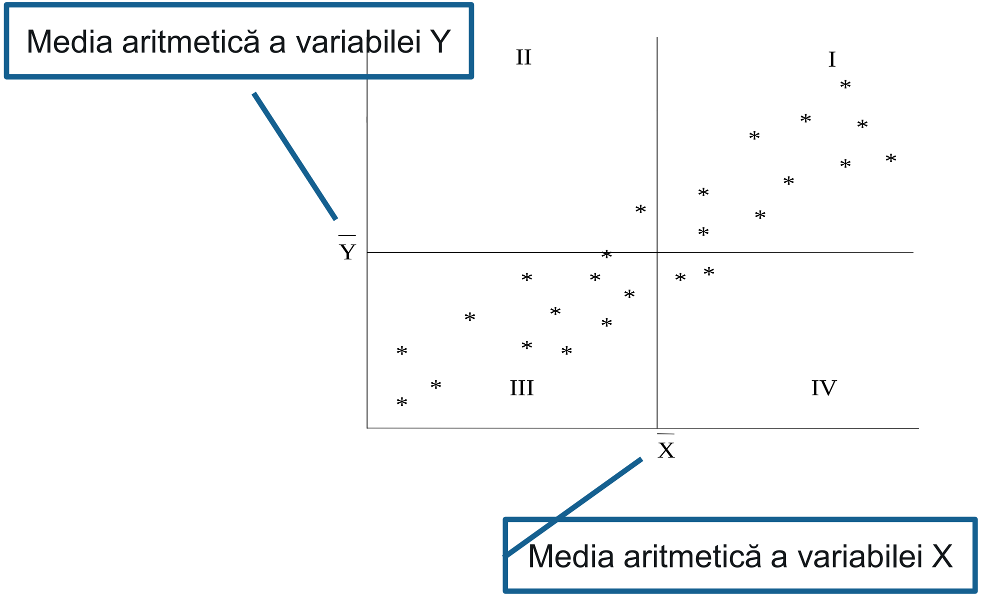


Diagrama de dispersie

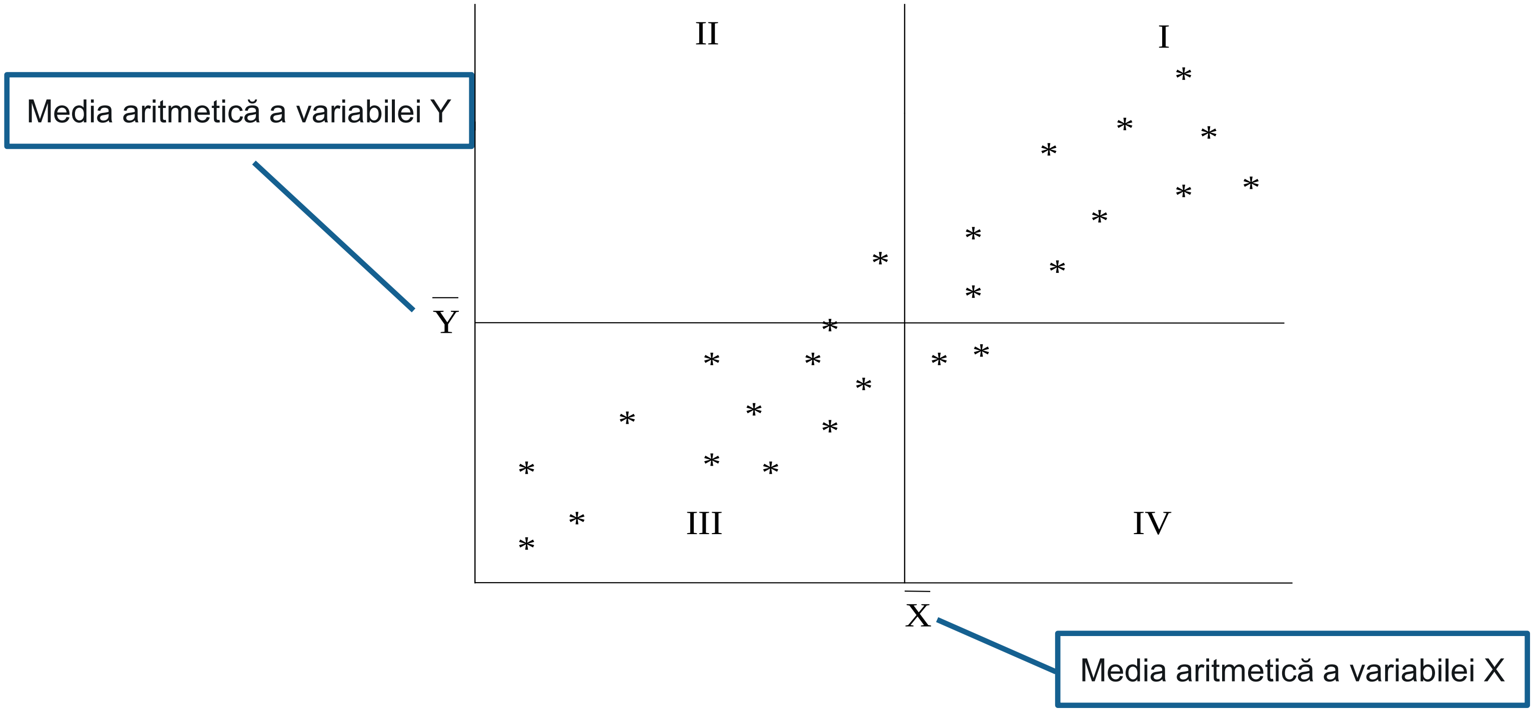


Diagrama de dispersie

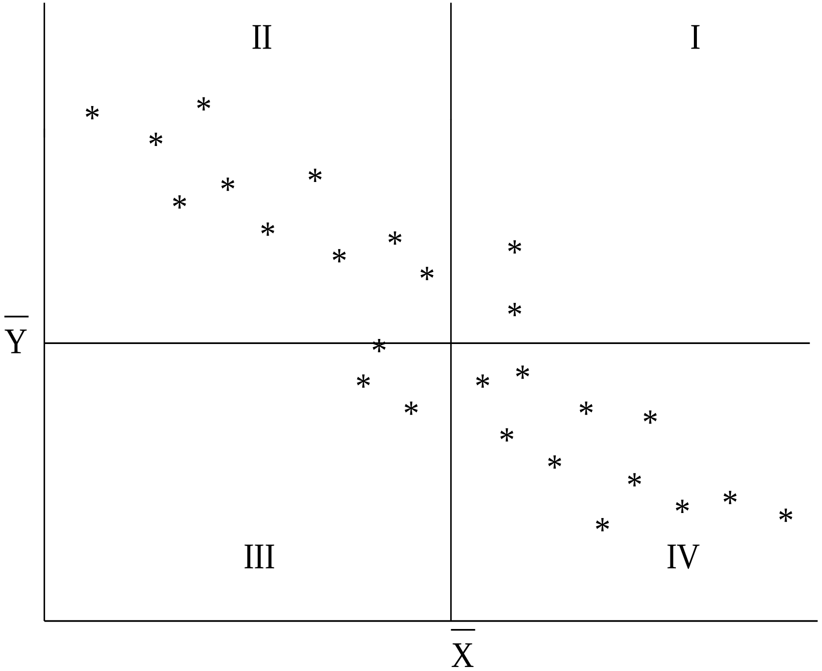
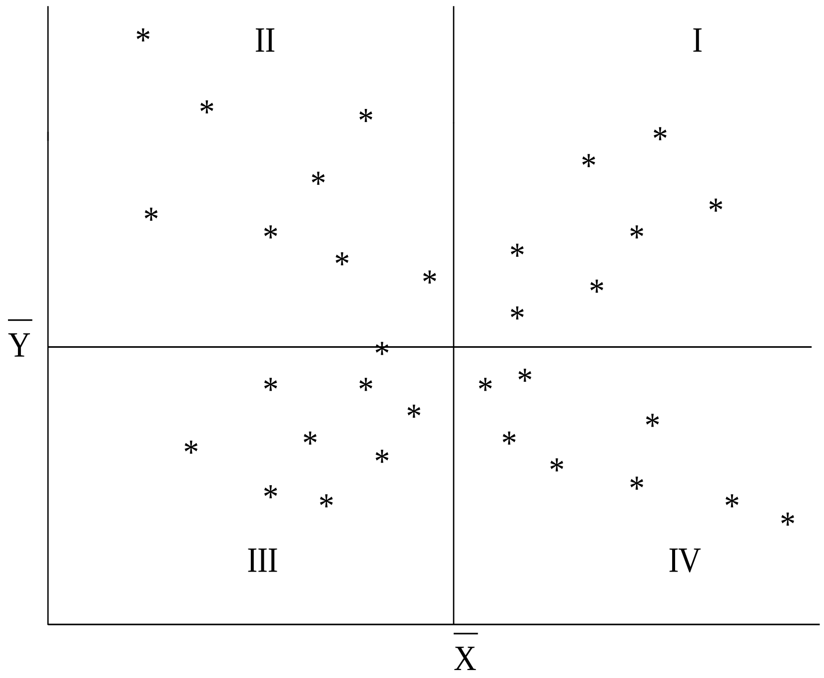


Diagrama de dispersie



Indici de corelaţie – Suma produselor ecart

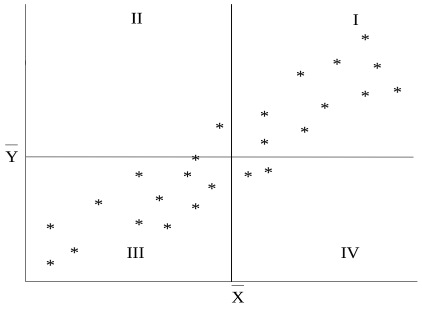
O măsură a intensităţii relaţiei dintre variabilele X şi Y este dată de suma:

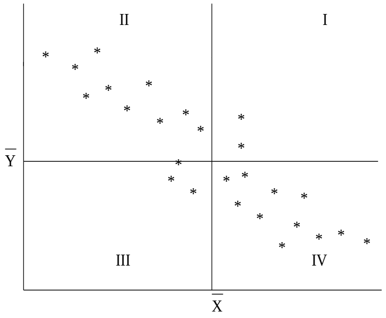
$$SPE = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Un dezavantaj evident al SPE este faptul că acest coeficient depinde de numărul de puncte din seria statistică şi de unităţile de măsură ale variabilelor

Indici de corelaţie – Suma produselor ecart

Pentru a descrie "intensitatea" relaţiei dintre cele două variabile X şi Y se utilizează observaţia că dacă punctul (Xi,Yi) se află în cadranele I sau III ale diagramei de dispersie atunci produsul este **pozitiv**, iar atunci când este situat în cadranele II şi IV este **negativ**.





SPE va fi cu atât mai mare în valoare absolută cu cât norul de puncte este mai apropiat de o alură generală crescătoare (SPE > 0) sau descrescătoare (SPE < 0).

Indici de corelaţie –Covarianţa COV(X,Y)

Pentru a obţine o mărime independentă faţă de volumul seriei statistice se utilizează covarianţa seriilor X şi Y, calculată prin:

$$COV(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})$$

Coeficientul de corelaţie Bravais-Pearson

Pentru a obţine un indicator independent şi de unităţile de măsură ale celor două variabile se utilizează coeficientul de corelaţie sau coeficientul Pearson:

$$r = \frac{COV(X,Y)}{Sx \cdot Sy} = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X^2\right)\right]\left[n\sum Y^2 - \left(\sum Y^2\right)\right]}}$$

S_X şi S_Y reprezintă abaterile standard pentru seriile X şi respectiv Y:

$$S = \sqrt{s^2}$$

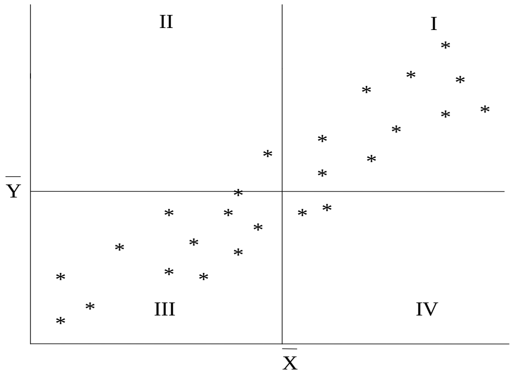
$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

Coeficientul de corelație (liniară) Pearson

- **Scop**: evaluarea asocierii a două variabile cantitative **(dintr-un eșantion)** din punctul de vedere al direcției și a importanței asocierii.
- Este o măsură a corelației liniare a celor două variabile cantitative care ne arată gradul de aproximare a punctelor la o linie care trece printre puncte (trendline).
- Este independent de unitatea de măsură a variabilelor
- Condiții de aplicare:
  - perechi de observații independente în eșantion
  - variabile cantitative
  - **ambele variabile au o distribuție normală**
  - relația dintre cele două variabile este liniară simplă (nu pătratică, exponențială ...)
  - Între cele două variabile trebuie să existe o **relație de cauzalitate**

Interpretarea coeficientului de corelaţie

- Coeficientul de corelaţie măsoară intensitatea relaţiei dintre variabilele X şi Y; $r \in [-1; 1]$



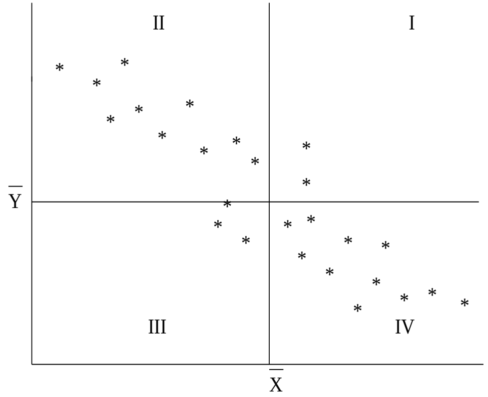
- Dacă r=1, punctele sunt situate pe o dreaptă de pantă pozitivă (crescătoare).

- Dacă 0 < r < 1, norul de puncte poate fi înlocuit (ajustat) printr-o dreaptă de pantă pozitivă.

- Când r este pozitiv relaţia între variabilele X şi Y este "pozitivă", adică o creştere a lui X determină în general o creştere a lui X.

Interpretarea coeficientului de corelaţie

- Dacă -1 < r < 0 atunci norul de puncte poate fi aproximat cu o dreaptă de pantă negativă. Dispersia punctelor faţă de dreaptă va fi cu atât mai mică cu cât r este mai apropiat de -1.

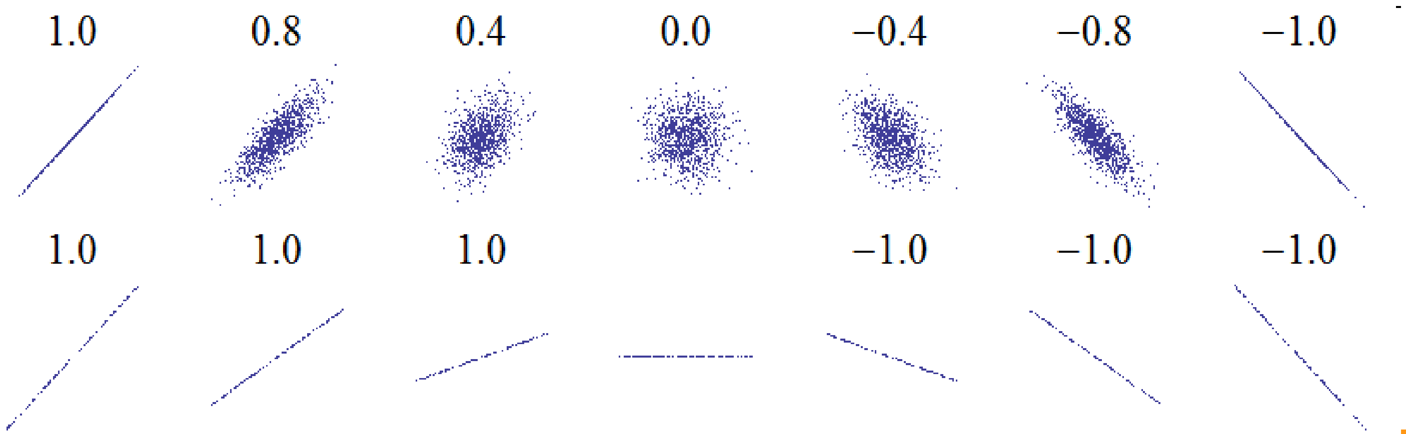


- Dacă r=-1 atunci toate punctele sunt situate pe o dreaptă de pantă negativă.

- Când r < 0 relaţia între cele două variabile este "negativă" adică o creştere a lui X are în general ca şi consecinţă o diminuare a lui Y.

Evaluarea graficului scatter (nor de puncte)

- În cazul în care relația este liniară, se poate evalua (subiectiv) puterea relației liniare. Cu cât punctele sunt mai apropiate de dreapta care le aproximează tendința, cu atât este mai puternică legătura; cu cât punctele sunt mai îndepărtate, cu atât legătura este mai slabă



Interpretarea Coeficientul de corelație Pearson (r) – direcția

- $r>0$ ➡ tendință de creștere / pantă ascendentă / legătură direct proporțională / corelație pozitivă
- $r<0$ ➡ tendință descrescătoare / pantă descendentă / legătură invers proporțională / corelație negative
- $r \cong 0$ ➡ nicio tendință
- cu cât este mai mare r (în valoare absolută), cu atât relația este mai puternică
- cu cât r este mai aproape de 0, relația este mai slabă

Interpretarea Coeficientul de corelație Pearson (r) – intensitatea

- Regulile empirice ale lui Colton (1974)

(-0.25; 0.25) => corelație slabă sau nulă

[0.25; 0.50) sau [-0.5; -0.25) => corelație acceptabilă

[0.50; 0.75) sau [-0.75; -0.5) => corelație moderată spre bună

[0.75 ; 1] sau [-1; -0.75] => corelație foarte bună

Interpretarea coeficientul de corelație Pearson (r) – testul de semnificație

- H0: Nu există o corelație liniară semnificativă statistic între variabilele x și y
- H1: Există o corelație liniară semnificativă statistic între variabilele x și y
- Decizia:
  - $p<0,05$=>respingem H0 și acceptăm H1; corelație liniară semnificativă statistic
  - $p>0,05$=>nu putem respinge H0; datele experimentale nu ne permit enunțarea existenței unei relații (la nivelul întregii populații) între variabilele luate în calcul; valoarea observată s-a datorat întâmplării

Interpretarea coeficientul de corelație Pearson (r și p)

Valoarea r	p > 0,05	p < 0,05
-0.25 la 0,25	corelație slabă sau nulă	corelație slabă sau nulă
0.25 la 0.50 (-0.5 la -0.25)	Nu are semnificatie statistica	un grad de asociere acceptabil
0.5 la 0.75 (-0.75 la -0.5)	Nu are semnificatie statistica	o corelație moderată spre bună
0.75 (sau mai mic decât -0.75)	Nu are semnificatie statistica	o foarte bună asociere sau corelație
r < -1; r > 1	Eroare	Eroare

Coeficientul de corelație - exemplu

- Coeficientul de corelație lineară Pearson pentru relația dintre trigliceride și greutate pentru 75 de subiecți este de **0,71**, iar valoarea p asociată este **0,004.** Perechile de observații sunt independente, **datele sunt distribuite normal, relația este liniară**

- Interpretarea valorii p asociate cu coeficientul de corelație:
  - p <0,05 => există o corelație liniară semnificativă statistic între trigliceride și greutate
  - Corelația liniară dintre trigliceride și greutate este semnificativă dpdv statistic
- Interpretarea direcției și puterii corelației:
  - relația este direct proporțională (r = 0,72> 0), iar puterea corelației este moderată până la bună (r = 0,72 - este în [0,50 și 0,75))

Coeficientul de corelație - exemplu

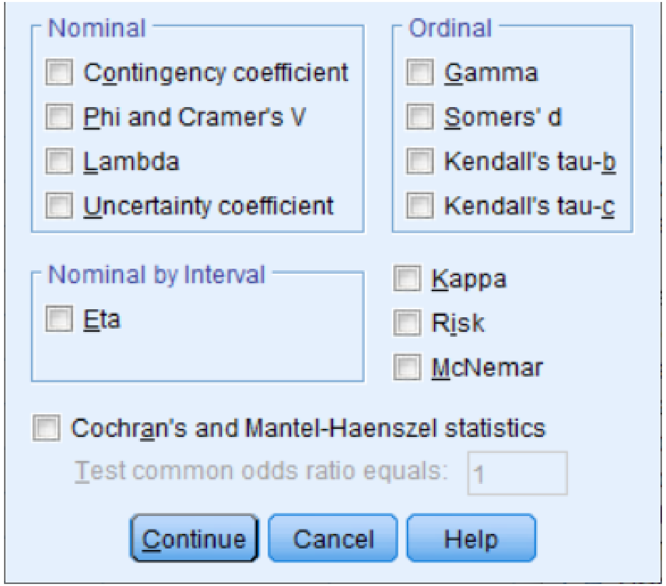
- Coeficientul de corelație lineară Pearson pentru relația dintre trigliceride și lungimea părului pentru 20 de subiecți este de **0,37**, iar valoarea p asociată este de **0,48**. Perechile de observații sunt independente, **datele sunt distribuite normal, relația este liniară**

- Interpretarea valorii p asociate cu coeficientul de corelație:
  - p> 0,05 => nu putem spune că există o corelație liniară semnificativă statistic între trigliceride și lungimea părului
- Interpretarea direcției și forței corelației:
  - Interpretarea nu este de niciun folos la nivelul populației, deoarece relația nu are semnificația statistică, rezultatul poate fi influențat de șansă.

Coeficientul de corelație Spearman (ρ–rho)

- **Scop**: evaluarea asocierii a două variabile cantitative din punctul de vedere al direcției și a importanței asocierii.
- Condiții de aplicare:
  - perechi de observații independente în eșantion
  - ambele variabile sunt ordinale/o variabilă ordinală și o cantitativă (cu distribuție normală sau nu)/ două variabile cantitative (cel puțin una nu are distribuție normală)
- Se folosesc rangurile seriei
- Utilitate: pentru a evalua relația dintre cele două variabile utilizate
- Interpretările sunt identice cu interpretările coeficientului de corelație Pearson

Alți coeficienți de corelație



- Coeficientul de determinare: $d=r^2$

COEFICIENTUL DE DETERMINARE

- este pătratul coeficientului de corelaţie, d = r*r.
- Prin definiţie, coeficientul de determinare reprezintă partea din variaţia totală a lui Y explicată prin relaţia liniară existentă între X şiY.
- Exprimat în procente, exprimă procentajul în care variaţia lui Y este dată prin relaţia liniară între cele două variabile.

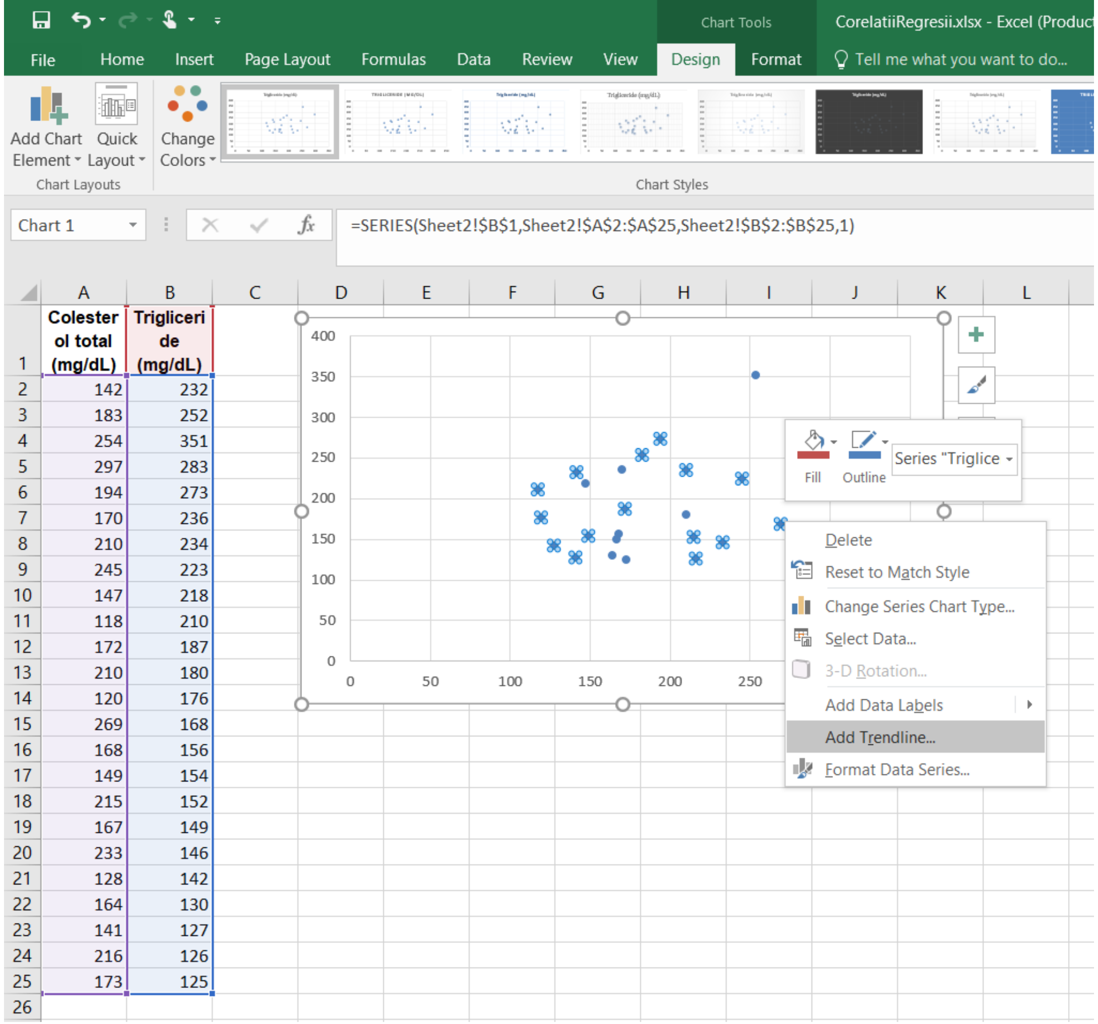
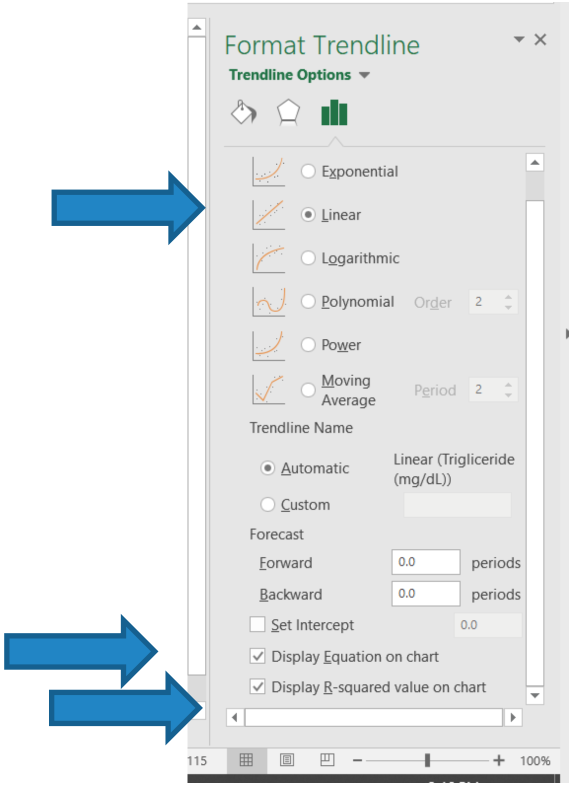
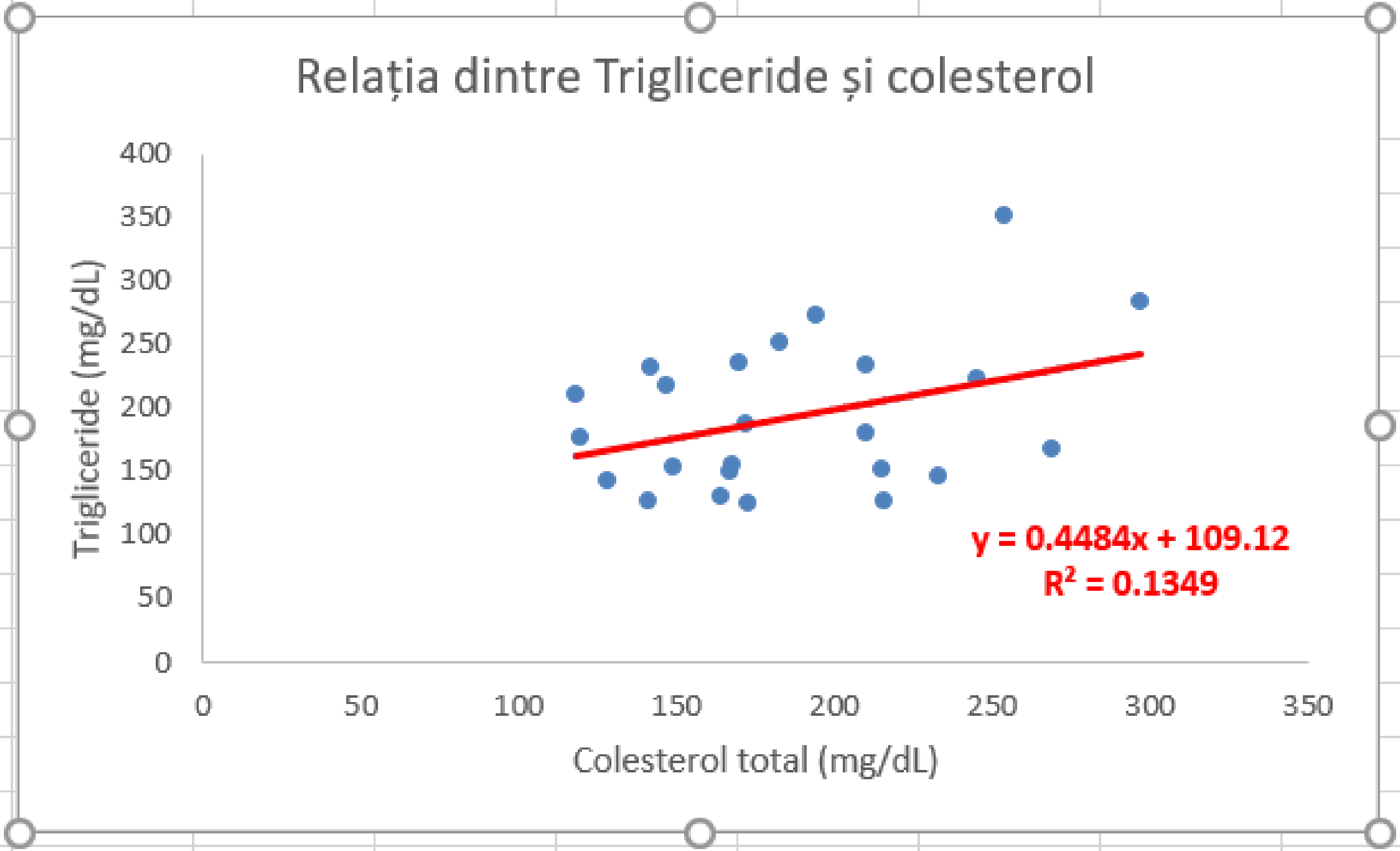


Chart Tools
CorelatiiRegresii.xlsx - Excel (Produc
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Design
Format
Tell me what you want to do...
Add Chart Element
Quick Layout
Change Colors
Chart Layouts
Chart Styles
Chart 1
=SERIES(Sheet2!$B$1,Sheet2!$A$2:$A$25,Sheet2!$B$2:$B$25,1)
Colesterol total (mg/dL)
Trigliceride (mg/dL)
142 232
183 252
254 351
297 283
194 273
170 236
210 234
245 223
147 218
118 210
172 187
210 180
120 176
269 168
168 156
149 154
215 152
167 149
233 146
128 142
164 130
141 127
216 126
173 125
Fill
Outline
Series "Triglice
Delete
Reset to Match Style
Change Series Chart Type...
Select Data...
3-D Rotation...
Add Data Labels
Add Trendline...
Format Data Series...

Format Trendline
Trendline Options
Exponential
Linear
Logarithmic
Polynomial
Order 2
Power
Moving Average
Period 2
Trendline Name
Automatic
Linear (Trigliceride (mg/dL))
Custom
Forecast
Forward 0.0 periods
Backward 0.0 periods
Set Intercept 0.0
Display Equation on chart
Display R-squared value on chart
100%



- 13% din variaţia trigliceridelor este dată de relaţia liniară dintre trigliceride şi colesterol

Regresia

- Metodă care studiază relația dintre 2 sau mai multe variabile
- O variabila dependentă
- Una sau mai multe variabile independente

Regresia

- Tipul de regresie:
  - În funcție de tipul variabilei dependente
    - variabilă cantitativă - regresie liniară
    - variabila dicotomială - regresie logistică
  - liniaritatea funcției
    - regresie liniară
    - regresie neliniară
  - numărul de variabile dependente:
    - regresie univariată (o variabilă dependentă)
    - regresie multivariată (2 sau mai multe variabile dependente)
  - numărul de variabile independente:
    - regresie simplă (o variabilă independentă)
    - regresie multiplă (2 sau mai multe variabile independente)

Regresia liniară simplă

- Scop: de a studia relația liniară dintre două variabile, unde variabila dependentă este cantitativă
- Obiective:
  - prezice valorile unei variabile cantitative în funcție de valorile celeilalte variabile
  - evaluează
    - existența unei asocieri / legături / relații între cele două variabile
    - direcția asocierii / legăturii / relației dintre două variabile
    - importanța asocierii / legăturii / relației dintre două variabile
- Condiții de aplicare:
  - variabilă dependentă - cantitativă
  - observații independente
  - existența unei relații liniare între cele două variabile (X și Y)
  - Reziduurile (diferentele dintre valorile observate si valorile estimate) urmeaza o distributie normala de medie 0
  - Reziduurile au aceeași variabilitate (varianță constantă)

Regresia liniară simplă

$$Y = a + bX + \varepsilon \quad (a, b = \text{coeficienții regresiei})$$

Calculul coeficienților a, b

Metoda celor mai mici pătrate - minimizăm suma tuturor pătratelor distanței fiecărui punct la linie

$$\sum_{i=1}^{n} \left(Y_i^{estimat} - Y_i\right)^2 \quad Y_i^{estimat} = a + bX_i$$

-obținem:

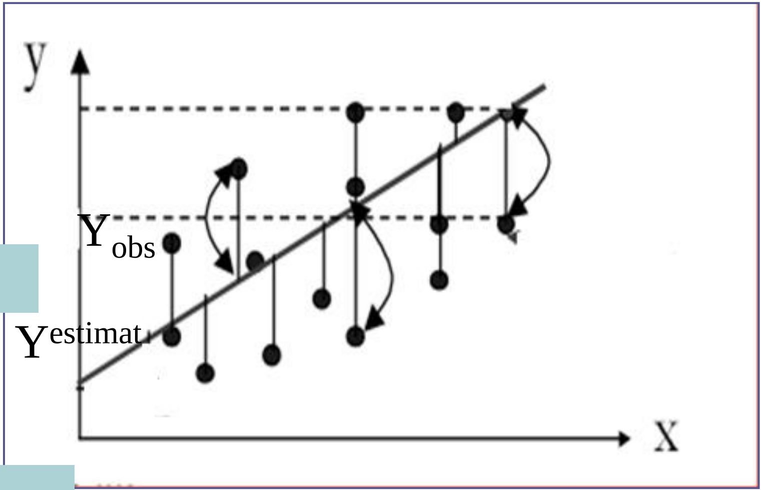
$$b = \frac{COV\ (X, Y)}{S_X^2}$$

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

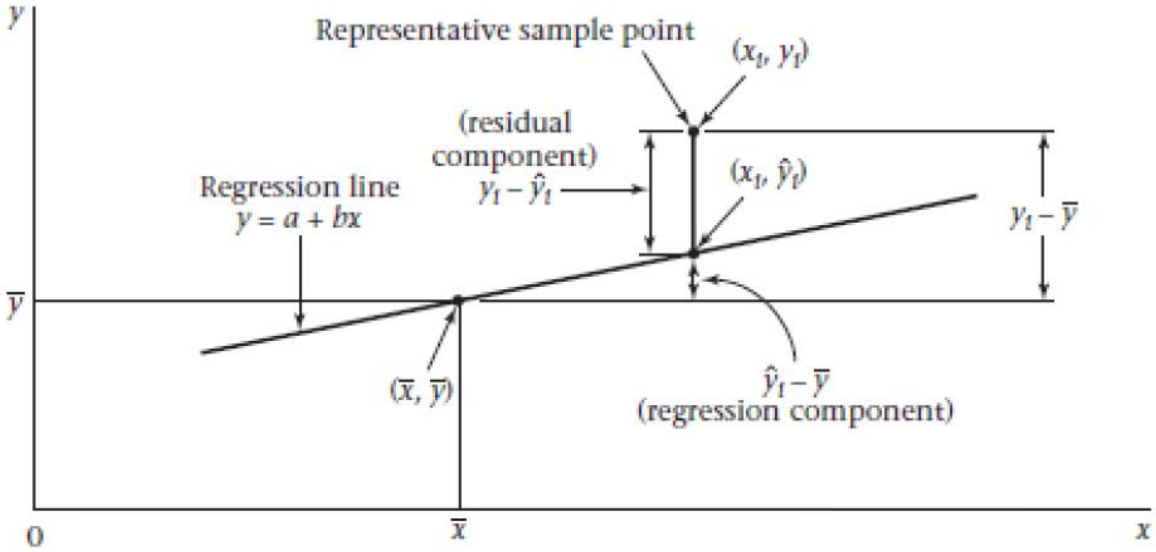
Valoarea observată

Covarianța

media lui Y, media lui X



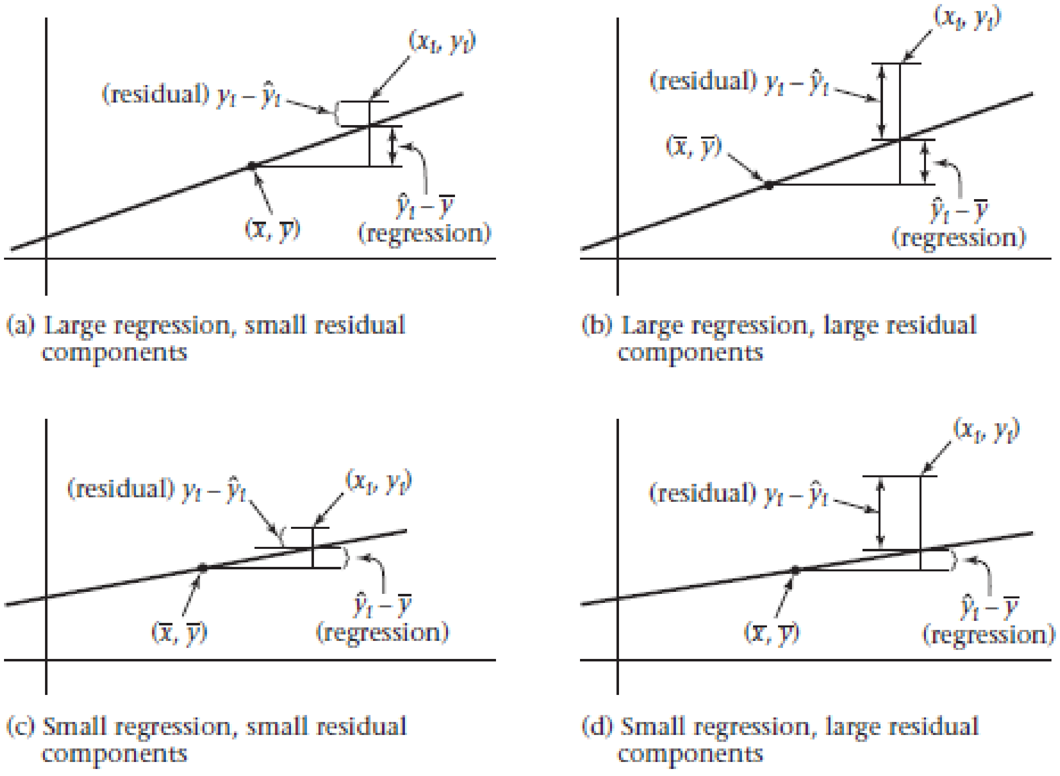
Distanța reziduală



Orice punct care nu se află pe dreapta de regresie, este la o distanță de aceasta numită **distanță reziduală**

Componenta reziduală și componenta de regresie

- O regresie “bună” are o componentă de regresie în general mai mare decât cea reziduală



(a) Large regression, small residual components

(b) Large regression, large residual components

(c) Small regression, small residual components

(d) Small regression, large residual components

Regresia liniară simplă - interpretare

- Dreapta de regresie Y (X): $Y = a + bX$
  - a = este valoarea lui Y când X este 0
  - b = panta liniei de regresie = pentru fiecare creștere cu o unitate a variabilei independente X, variabila independentă Y crește în medie cu valoarea coeficientului b
- b> 0 → tendință crescătoare / panta ascendentă / panta pozitivă / legătură direct proporțională
- b< 0 → tendință descrescătoare / panta descendentă / panta negativă / legătură invers proporțională
- b $\cong$ 0 → fără tendință
- cu cât b este mai mare (în valoare absolută), cu atât relația este mai puternică
- cu cât b este mai aproape de 0, cu atât relația este mai slabă

Regresia lineară multiplă

- Scop: de a studia relația liniară dintre mai mult de două variabile și o variabilă dependentă cantitativă
- Obiective:
  - prezice valorile unei variabile cantitative pe baza valorilor altor variabile
  - evaluează
    - existența unei asociații / legături / relații între cele două variabile
    - direcția asocierii / legăturii / relației dintre două variabile
    - importanța asocierii / legăturii / relației dintre două variabile
  - ajustare / corectare - efectul altor variabile
- Condiții de aplicare:
  - observațiile sunt independente una de cealaltă
  - relația dintre fiecare predictor si variabila dependenta este liniară

Regresia lineară multiplă – cuantificarea importanței relației între mai multe variabile

- Relațiile dintre diferite variabile medicale sunt adesea complexe și sunt implicați mai mulți factori
- Folosind coeficientul de corelație sau cu regresie liniară simplă putem evalua relația dintre **două** variabile **-** o **analiză univariază** (între o **variabilă dependentă -** și o **variabilă explicativă / independentă**)
- Este necesar să luăm în considerare efectul altor variabile (numite confundante sau alte variabile explicative) despre care se știe că influențează relația pe care o studiem. Trebuie să facem un studiu bibliografic pentru a identifica aceste variabile de confundare
- Pentru a lua în considerare mai multe variabile, trebuie să se utilizeze tehnici multivariate (de exemplu, regresie liniară multiplă, regresie logistică multiplă, ……)

Regresie liniară multiplă - Cuantificarea importanței relației pentru mai multe variabile

- Dacă există mai multe variabile care sunt legate de variabila dependentă de interes, importanța lor poate fi evaluată cu o **regresie liniară multiplă,** care oferă un coeficient (numit **ajustat (adjusted)**) pentru fiecare variabilă independentă:
- Variabila dependentă (prezisă, explicată) a regresiei:
  - Trigliceride (mg / dL)
- Variabile independente (explicative, predictive)
  - Calitativ (sex)
  - Cantitativ (greutate (kg))
- Coeficientul **ajustat (standardizat)** ne poate aduce mai aproape de adevăr decât coeficientul **brut** (**crude**, **unadjusted, unstandardized**) a unei regresii liniare simple sau a unui coeficient de corelație simplu, deoarece putem lua în considerare alte variabile care acționează în același timp

Regresie liniară multiplă - Cuantificarea importanței relației pentru mai multe variabile

- Ecuația de regresie multiplă liniară:
- Variabilă dependentă = $coef_1 * var_1 + coef_2 * var_2 + ... + coef_n * var_n + coef_0$
- Ex: trigliceride (mg / dL) = 23,10 * obezitate (da/nu) + 1,14 * colesterol (mg/dL)

- Interpretarea **coeficientului ajustat** pentru variabilele calitative **dicotomice** (obezitate):
- Trigliceridele cresc în medie cu 23,1 mg / dL pentru cei care sunt obezi (obezitate=da) comparativ cu cei care nu sunt obezi (obezitate=nu), dacă păstrăm celelalte variabile constante (colesterolul)

Regresie liniară multiplă - Cuantificarea importanței relației pentru mai multe variabile

- Ecuația de regresie multiplă liniară:
- Variabilă dependentă = $coef_1 * var_1 + coef_2 * var_2 + \ldots + coef_n * var_n + coef_0$
- Ex: trigliceride (mg / dL) = 23,10 * obezitate (da/nu) + 1,14 * colesterol (mg/dL)

- IInterpretarea **coeficientului ajustat** pentru variabilele **cantitative** (colesterol):
- Trigliceridele cresc în medie cu 1,14 mg/dL pentru fiecare un mg/dL de colesterol, dacă menținem celelalte variabile constante (obezitate)

Interpretarea regresiei multiple

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound	Correlations Zero-order	Correlations Partial	Correlations Part	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	25.838	2.882		8.964	.000	20.090	31.585					
	People who read (%)	.315	.034	.636	9.202	.000	.247	.383	.869	.738	.465	.535	1.868
	Daily calorie intake	.007	.001	.342	4.949	.000	.004	.010	.776	.506	.250	.535	1.868

a. Dependent Variable: Average female life expectancy

- Valorile brute, raw (nestandardizate) si valorile standardizate ale ponderilor regresiilor pentru *speranța de viață a femeilor* în funcție de *aportul zilnic de calorii* și de *procentul celor care citesc din anturaj*.
- Valorile standardizate ale coeficientilor de regresie (beta weight) pentru *daily caloric intake* este **0.342**, iar pentru *percentage of people who read* este mult mai mare: **0.636**.
- Interpretare: pentru fiecare unitate *din percentage of people who read*, variabilaY (*female life expectancy*) va crește cu un multiplude **0.636**. Amândoi coeficienții de corelație sunt semnificativi statistic, p < .001

Regresia logistică

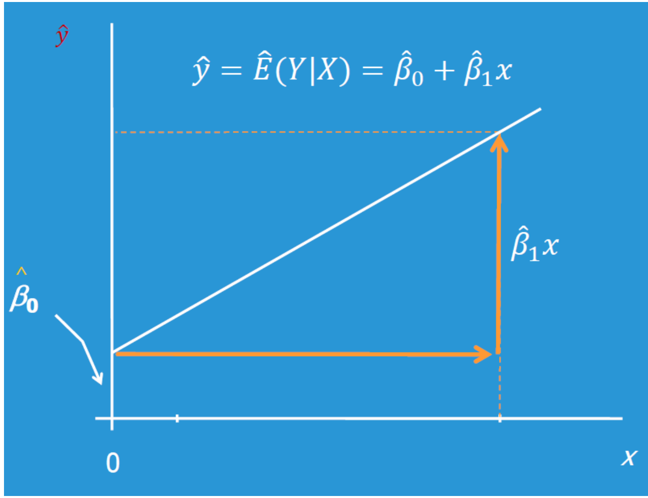
- Tipuri de variabile într-o regresie logistică:
  - Variabila dependentă (variabila pe care încercăm să o prezicem) trebuie să fie **o variabilă dicotomială**!
    - Prezența bolii (sau decesului)
      - diabet, hipertensiune (da/nu)
  - Variabile independente (variabile cu care se încearcă prezicerea variabilei dependente) - **una sau mai multe, de diferite tipuri**
    - Calitative (consumul de alcool, fumatul, obezitatea)
    - Cantitative (vârsta)
- Conditii de aplicabilitate
  - observațiile sunt independente una de cealaltă
  - relația dintre fiecare predictor continuu si variabila dependenta este liniară pe scala logit.

Regresia logistică

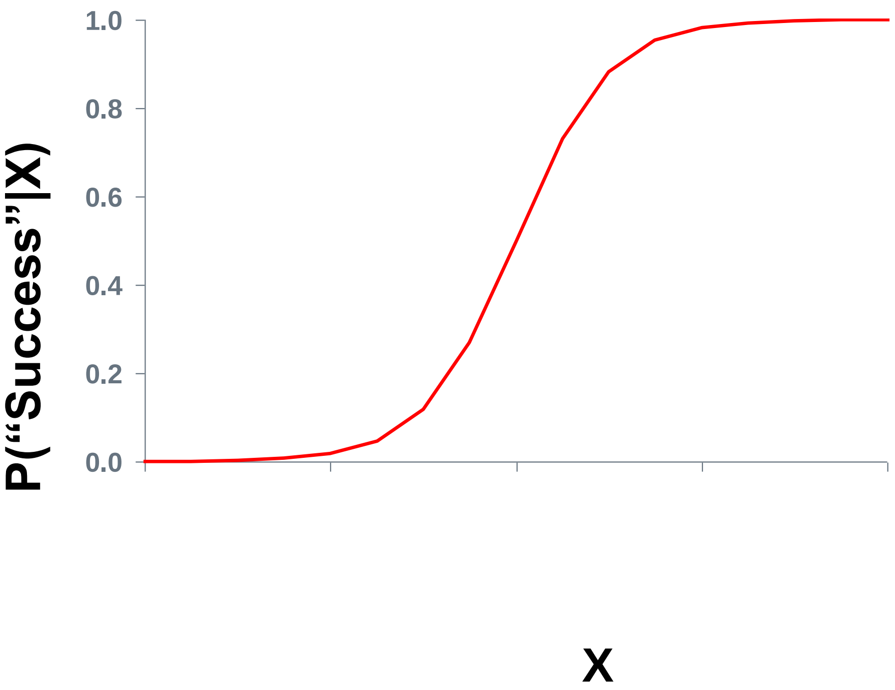
- Ne permite predicția unei variabile discrete printr-un mix de predictori, variabile atât continue cât și discrete
- Răspunde la aceeași întrebare: există o funcție de discriminare dar nu este necesară existența unor condiții pentru predictori: normalitate, relație liniară...
- Clasificarea poate să fie doar discretă
- Transformarea logit

$$\ln(\text{șansă}) = \ln\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 X$$

$$p = \frac{e^{\beta_o+\beta_1 X}}{1+e^{\beta_o+\beta_1 X}}$$



Funcția logistică



Regresia logistică

▷ **Modelul de regresie logistică:**

$$ln(\frac{p}{1-p}) = b_0 + b_1x_1 + b_2\, x_2 + b_3x_3 + \cdots + b_kx_k$$

- p - probabilitatea ca un subiect cu un set de valori ale variabilelor independente să aibă evenimentul de interes (boala sau un alt rezultat urmărit)
- $X_1, X_2, \ldots, X_k$ sunt variabile independente
- $ln(\frac{p}{1-p}) = logit(p)$
- b_0 reprezintă termenul liber (constanta modelului)
- $b_1, \ldots, b_k$ sunt coeficienți de regresie parțială. Atunci când se transformă de pe scala logartimică la scala naturală, devin OR-uri (odds ratio), deci exponențiala unui coeficient al regresiei logistice reprezintă raportul sanselor (OR) pentru creștere cu o unitate a valorii variabilei independente

- Exemple...

EXEMPLU
POATE FI PREZIS SUCCESUL UNUI STUDENT?

- Variabile predictori ai succesului
- Lot: 315 studenti
- Obiectiv (variabilă dependentă):
  - Succes (1 –DA, 0 -NU).
- Predictori:
  - Date cantitative: Markeri ai inteligenței (IQ) și ai pregătirii studentului (medii anterioare în facultate)
  - Date calitative: sex

POATE FI PREZIS SUCCESUL UNUI STUDENT FOLOSIND PREDICTORI MULTIPLI?

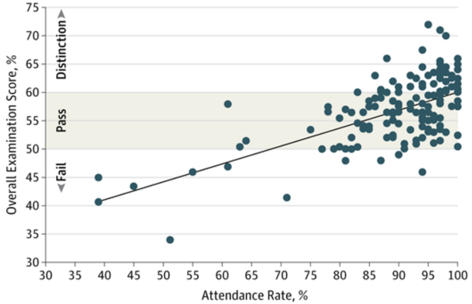
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Sex(1)	1,431	,340	17,677	1	,000	4,185	2,147	8,156
	Coeficient de inteligenta	,144	,019	58,247	1	,000	1,155	1,113	1,198
	Medie in facultate	,556	,121	21,118	1	,000	1,743	1,375	2,209
	Constant	-18,046	2,046	77,823	1	,000	,000		

a. Variable(s) entered on step 1: Sex, Coeficient de inteligenta, Medie in facultate.

- ln(șansă)=-18,046+1,431(sex)+0,144*IQ+0,556*Medie
- Șansa=$e^{[18,046+1,431(sex)+0,144*IQ+0,556*Medie]}$

Exemplu: Corelatie



Student Attendance and Academic Performance in Undergraduate Obstetrics/Gynecology Clinical Rotations FREE

Richard P. Deane, MB BCh[1]; Deirdre J. Murphy, MD[1] *JAMA*. 2013;310(21):2282-2288. doi:10.1001/jama.2013.282228.

http://journalgateway.com/ijomp/article/view/583

Objective To evaluate the relationship between student attendance and academic performance in a medical student obstetrics/gynecology clinical rotation.

Design, Setting, and Participants A prospective cohort study of student attendance at clinical and tutorial-based activities during a full academic year (September 2011 to June 2012) within a publicly funded university teaching hospital in Dublin, Ireland. Students were expected to attend 64 activities (26 clinical activities and 38 tutorial-based activities) but attendance was not mandatory. All 147 fourth-year medical students who completed an 8-week obstetrics/gynecology rotation were included.

Exposures Student attendance at clinical and tutorial-based activities, recorded using a paper-based logbook.

Main Outcomes and Measures The overall examination score (out of a possible 200 points) was obtained using an 11-station objective structured clinical examination (40 points), an end-of-year written examination comprising 50 multiple-choice questions (40 points) and 6 short-answer questions (40 points), and an end-of-year long-case clinical/oral examination (80 points). Students were required to have an overall score of 100 points (50%) and a minimum of 40 points in the long-case clinical/oral examination (50%) to pass.

Results The mean attendance rate was 89% (range, 39%-100% [SD, 11%], n = 57/64 activities). Male students (84% attendance, $P = .001$) and students who failed an end-of-year examination previously (84% attendance, $P = .04$) had significantly lower rates. **There was a positive correlation between attendance and overall examination score ($r = 0.59$ [95% CI, 0.44-0.70]; $P < .001$).** Both clinical attendance ($r = 0.50$ [95% CI, 0.32-0.64]; $P < .001$) and tutorial-based attendance ($r = 0.57$ [95% CI, 0.40-0.70]; $P < .001$) were positively correlated with overall examination score. The associations persisted after controlling for confounding factors of student sex, age, country of origin, previous failure in an end-of-year examination, and the timing of the rotation during the academic year. Distinction grades (overall score of ≥60%) were present only among students with attendance rates of 80% or higher. The odds of a distinction grade increased with each 10% increase in attendance (adjusted odds ratio, 5.52; 95% CI, 2.17-14.00). The majority of failure grades (6/10 students; 60%) occurred in students with attendance rates lower than 80%. The adjusted odds ratio for failure with attendance rates of 80% or higher was 0.11 (95% CI, 0.02-0.72)..

- Multumesc pentru atenție!