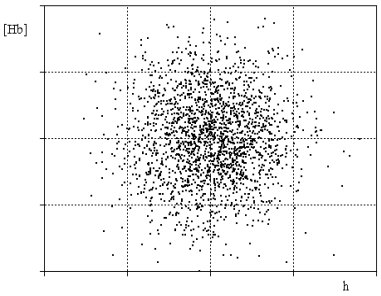
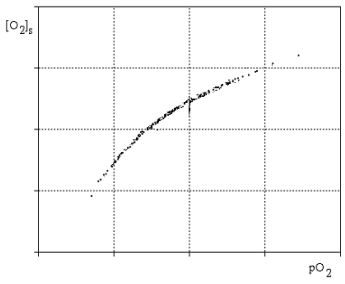
Analiza corelației și regresiei

1. Aspecte teoretice

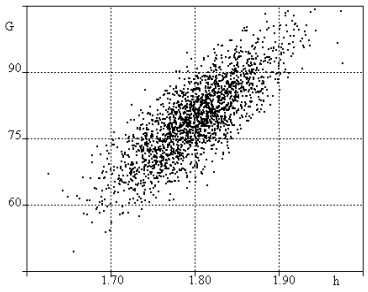
a. **Definiție:** *studiază dependenţa dintre două sau mai multe variabile.*
b. **Regresia:** *ne arată cum o variabilă este dependentă de altă variabilă* (una dintre variabile este fixată și cealaltă aleatorie)
   Obs: presupune existența unei relații cu sens (o variabilă o influențează pe cealaltă)
c. **Corelaţia:** *ne arată gradul în care o variabilă este dependentă de o altă variabilă* (măsoară tăria asociației dintre variabile prin calculul coeficientului de corelație, ambele variabile sunt aleatorii)
d. Diagrame de corelație: reprezentări grafice având pe axe cele două variabile (grafic de tip *Scatter*)
e. **Analiza bivarată:** *analiza care urmarește comportamentul a 2 variabile*
   i. Variabila X – predictor
   ii. Variabile Y – răspuns
f. Relatii de dependenta - repartitia punctelor este aproape simetrica si uniforma (un individ = un punct)
   i. Variabile independentre
   ii. Variabile dependente
   iii. Variabile corelate



Variabile independente

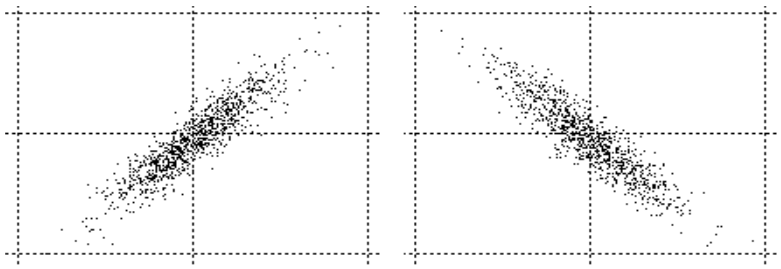


Variabile dependente



Variabile corelate

g. **Semnificația coeficientului de corelație:** se testează cu testul **t** și se interpretează ca orice test (dacă $p > 0.05$, corelația este nesemnificativă ; dacă $p < 0.05$, corelația este semnificativă)
h. **Corelația liniară, coeficientul de corelație Pearson:** caracterizează "intensitatea" asociației dintre 2 variabile
   Proprietăți: $r \in [-1, 1]$
   $r > 0 \Rightarrow$ corelație directă (pozitivă)
   $r < 0 \Rightarrow$ corelație indirectă (inversa, negativă)
   $r = 1 \Rightarrow$ corelație directă perfectă
   $r = -1 \Rightarrow$ corelație inversă perfectă
   $r = 0 \Rightarrow$ variabile necorelate



a. Corelație directă b. Corelație inversă

Colton (1974) sugerează următoarele reguli empirice privind interpretarea coeficientului de corelație:

r ∈ [-0.25, 0.25] înseamnă o **corelație slabă** sau **nulă (cand se apropie de 0)**

r ∈ [0.25, 0.50] sau **[-0.50, -0.25]** înseamnă o **corelatie slaba spre medie**

r ∈ [0.5, 0.75] sau **[-0.75, -0.5]** înseamnă o **corelatie medie spre buna**

r > 0,75 sau r < -0,75 înseamnă o **corelatie foarte bună** sau **puternica (cand se apropie de ±1)**

Coeficientul de determinare R^2: *reprezintă proporția din variația uneia dintre variabile ce poate fi atribuită (sau explicată) de variația celeilalte variabile (oferă informații despre tăria corelației, nu despre sensul corelației)*

Proprietăți: $R^2 \in [0\,,1]$

Semnificația statistică: se testează dacă apariția corelației este întâmplătoare sau e reproductibilă în populație.

Notații: r – pentru eșantion și ρ – pentru populație

Ipotezele statistice:

H_0: $\rho = 0$, (cele două variabile sunt independente)

H_a: $\rho \neq 0$, (cele două variabile sunt corelate)

Testul aplicat:

Testul **t** (Student), are o repartiție Student cu N-2 grade de libertate

$$t = t_{calc} = r \cdot \sqrt{\frac{N-2}{1-r^2}}$$

Dreapta de regresia: *"cea mai bună" dreaptă, care trece prin punctele diagramei de corelație*

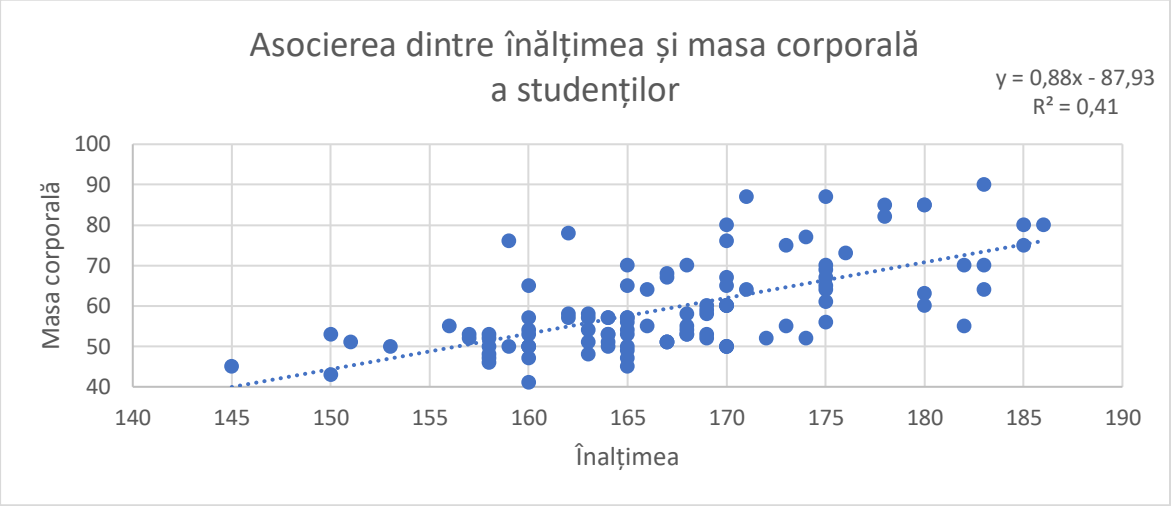
Parametrii dreptei: $y = a + bx$, unde a este ordonata la origine (intercept), iar b este panta (slope)

Tehnica de fitare – metoda celor mai mici pătrate (Gauss). Pentru determinarea coeficienților a și b din ecuația dreptei vom considera dreapta care trece printre punctele experimentale ca având suma pătratelor abaterilor minimă

$$SSE = \sum \epsilon_i^2 = min$$

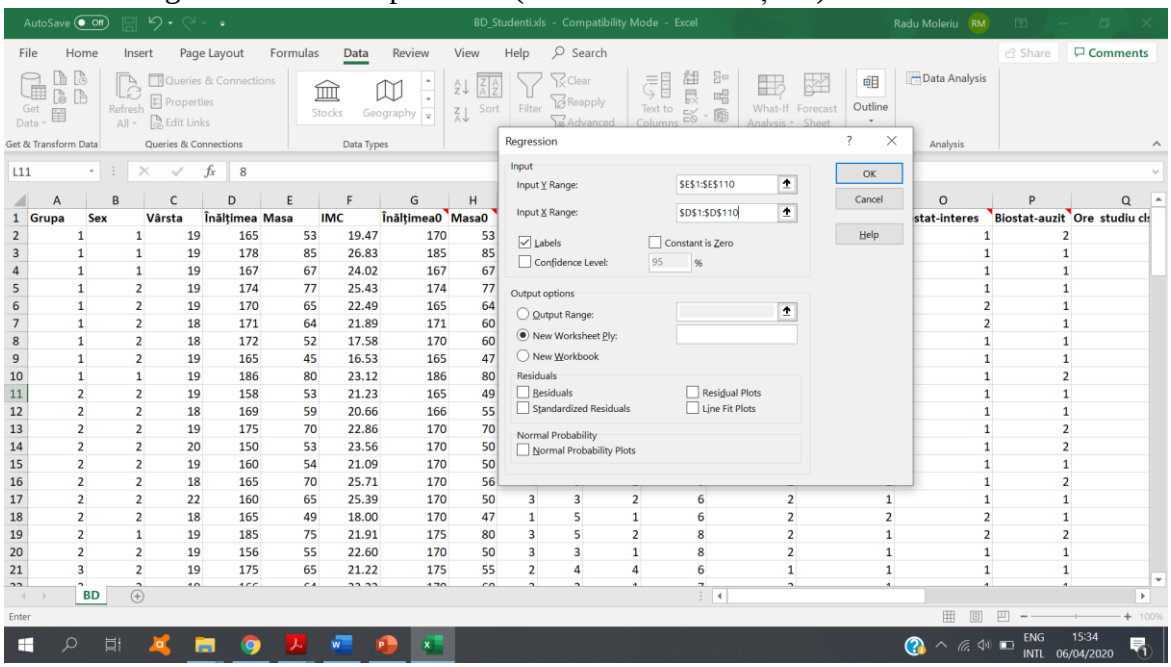
2. Model exercițiu rezolvat

▶ Ex1: Setul de date conține înălțimile și masa corporală a studenților de la Facultatea de Farmacie, anul I (109 studenți). Există o asociere între înălțimea și masa corporală a studenților?



Data->Data Analysis->Regression

Input X range -> variabila independenta (ex-> variabila masa corporală)
Y range -> variabila dependenta (ex-> variabila înălțime)



Rezultatul generat de Microsoft Excel

Regression Statistics					
Multiple R	0,64				
R Square	0,41				
Adjusted R Square	0,41				

Standard Error	8,64				
Observations	109				
ANOVA					
	df	SS	MS	F	Significance F
Regression	1	5614,29	5614,29	75,19	5,10E-14
Residual	107	7989,81	74,67		
Total	108	13604,11			
	Coefficients	Standard Error	t Stat	P-value	
Intercept	-87,93	17,04	-5,16	1,15E-06	
X Variable 1	0,88	0,10	8,67	5,10E-14	

Ipoteza de cercetare:

În acest studiu dorim să vedem dacă există o asociere între înălțimea studenților și masa corporală a studenților.

Ipoteze statistice:

H0: cele 2 variabile sunt independente (sau – nu există o asociere semnificativă între variabilele testate)

$\rho = 0$

Ha: cele 2 variabile sunt dependente (sau –există o asociere semnificativă între variabilele testate)

$\rho \neq 0$ (ipoteză alternativă bilaterală)

Alegerea testului:

Deoarece avem 2 variabile numerice independente și dorim să testăm gradul de asociere dintre ele vom aplica analiza de corelație și regresie.

Data->Data Analysis->Regression

Input X range -> variabila independenta (ex-> variabila înălțime)

Y range -> variabila dependenta (ex-> variabila masa corporală)

Interpretarea rezultatelor:

r - coeficentul lui Pearson

$r = 0.64$, $r > 0$ => corelație directă, pozitivă

$r \in [0.5; 0.75]$ => corelație medie ca intensitate

R^2 – coeficent de determinare

$R^2 = 0.41$ => în procent de 41% masa corporală a studenților este influențată de înălțimea acestora

$p = 5.10E - 14 = 5.10 * 10^{-14} = 0.0000000000000510$ =>

$p < 0.05, p < 0.01, p < 0.001$ => respingem H0 => variabilele sunt extrem de corelate

Concluzia statistică:

$p < 0.05, p < 0.01, p < 0.001$ => respingem H0 => variabilele sunt extrem de corelate

Concluzia de cercetare:

În urma aplicării analizei de corelație și regresie putem spune că există o asociere pozitivă, directă, medie ca intensitate și extrem de semnificativă din punct de vedere statistic

3. Probleme propuse (pentru urmatoarele prelucrări de date se va folosi fisierul BD_Studenti.xls)

Problema 1:

Investigati posibila asociere intre masa corporala a unei persoane (exprimata de variabila Greutate) si inaltimea aceleiasi persoane (exprimata de variabila Inaltime).

1. Ce tip de investigatie statistica este aceasta? Motivati-va afirmatia si explicati care sunt caracteristicile.
2. Formulati ipotezele statistice in termeni matematici si in cuvinte.
3. Reprezentati grafic datele. Alegeti cea mai potrivita reprezentare grafica, asfel incat ea sa fie in concordanta cu afirmatiile facute la punctele 1 si 2.
4. Prelucrati datele din punct de vedere statistic, potrivit afirmatiilor pe care le-ati facut la punctele anterioare.
5. In ce proportie variatia greutatii unui individ este cauzata de alti factori decat variatia inaltimii.
6. Interpretati rezultatele obtinute. Formulati atat in termeni statistici, cat si interpretând termenii statistici in relatie cu reprezentarea grafica si cu ipotezele formulate initial.
7. Bazandu-ne pe aceasta analiza ce greutate ar trebui sa aiba un individ care are o inaltime de 1.78 m? Dar un individ care are 2.35 m?

Problema 2:

Investigati posibila asociere intre indicele de masa corporala (IMC=m(KG)/$h^2(m^2)$ a unei persoane si indicele de masa corporala dorit (calculati IMC0).

1. Ce tip de investigatie statistica este aceasta? Motivati-va afirmatia si explicati care sunt caracteristicile.
2. Formulati ipotezele statistice in termeni matematici si in cuvinte.
3. Reprezentati grafic datele. Alegeti cea mai potrivita reprezentare grafica, asfel incat ea sa fie in concordanta cu afirmatiile facute la punctele 1 si 2.
4. Prelucrati datele din punct de vedere statistic, potrivit afirmatiilor pe care le-ati facut la punctele anterioare.
5. Interpretati rezultatele obtinute. Formulati atat in termeni statistici, cat si interpretând termenii statistici in relatie cu reprezentarea grafica si cu ipotezele formulate initial.

Problema 3:

Investigati posibila asociere între numarul de ore de studiu in timpul examenelor (exprimata de variabila Ore-studiu anul I) și numarul orelor de somn (exprimata de variabila Ore - somn).

1. Ce tip de investigatie statistica este aceasta? Motivati-va afirmatia si explicati care sunt caracteristicile.
2. Formulati ipotezele statistice in termeni matematici si in cuvinte.
3. Reprezentati grafic datele. Alegeti cea mai potrivita reprezentare grafica, asfel incat ea sa fie in concordanta cu afirmatiile facute la punctele 1 si 2.
4. Prelucrati datele din punct de vedere statistic, potrivit afirmatiilor pe care le-ati facut la punctele anterioare.
5. Interpretati rezultatele obtinute. Formulati atat in termeni statistici, cat si interpretând termenii statistici in relatie cu reprezentarea grafica si cu ipotezele formulate initial.

Problema 4:

Investigati posibila asociere intre numarul de ore de studiu in timpul examenelor (exprimata de variabila Ore-studiu anul I) si numarul orelor de studiu din clasa a XII-a (exprimata de variabila Ore – studiu cls XII).

1. Ce tip de investigatie statistica este aceasta? Motivati-va afirmatia si explicati care sunt caracteristicile.
2. Formulati ipotezele statistice in termeni matematici si in cuvinte.
3. Reprezentati grafic datele. Alegeti cea mai potrivita reprezentare grafica, asfel incat ea sa fie in concordanta cu afirmatiile facute la punctele 1 si 2.
4. Prelucrati datele din punct de vedere statistic, potrivit afirmatiilor pe care le-ati facut la punctele anterioare.
5. Interpretati rezultatele obtinute. Formulati atat in termeni statistici, cat si interpretând termenii statistici in relatie cu reprezentarea grafica si cu ipotezele formulate initial.