



UNIVERSITATEA DE MEDICINĂ,  
FARMACIE, ȘTIINȚE ȘI TEHNOLOGIE  
„GEORGE EMIL PALADE”  
DIN TÂRGU MUREȘ

PROBABILITĂȚI ȘI STATISTICĂ ÎN SISTEME MEDICALE  
**CURSUL 8, 14-15 OCTOMBRIE 2020**

**ANALIZA REGRESIEI ȘI CORELAȚIEI PENTRU  
DATE DIN SISTEME MEDICALE**

**PARTEA II - ANALIZA REGRESIEI**

prof. univ. dr. habil Manuela Rozalia GABOR

# RELAȚII ÎNTRE VARIABLE CANTITATIVE

- Relația între două variabile cantitative poate fi abordată din două puncte de vedere

1

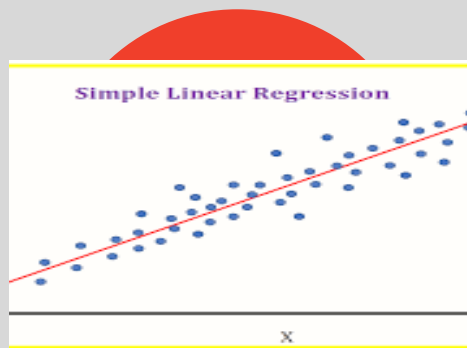
- Calculul numeric al intensității legăturii dintre cele două variabile se realizează cu ajutorul unor indici de corelație dintre care cel mai cunoscut este coeficientul de corelație

2

- Dacă acești indici denotă existența unei relații între variabile se poate determina tipul de legătură care face ca **Y** să depindă de **X**, adică determinarea unei funcții  $f$  numită funcție de regresie, astfel încât  $Y = f(X)$ .
- În acest caz, una dintre variabile, care se numește **variabila independentă (X)**, poate fi controlată nealeator, iar cealaltă numită **variabila dependentă (Y)** ia valori care nu sunt controlate (variază aleator).

- Cea mai simplă și cea mai utilizată funcție de regresie este cea liniară care aproximează și relația empirică de proporționalitate:  $f(X) = a + b X$ .

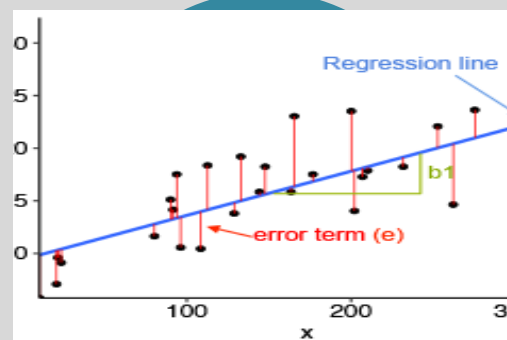
# UTILIZAREA FUNCTIILOR DE REGRESIE



Funcțiile de regresie sunt adesea utilizate pentru prezicerea (exprimarea) unei valori a lui Y pentru o valoare a lui X și invers.

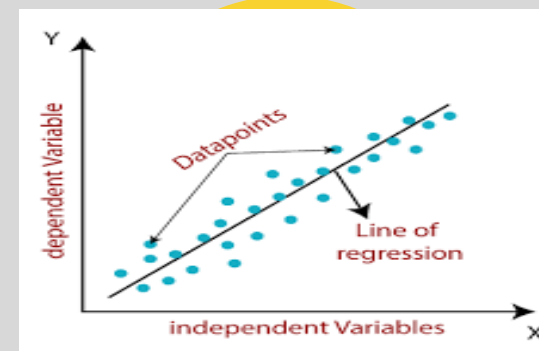
Să presupunem că funcția de regresie e de tipul  $y(x)$ . Când se determină valoarea funcției (adică a lui Y), pentru un X cuprins în intervalul  $[x_{min}, x_{max}]$ , atunci se efectuează o

Operație de interpolare, iar când X se află în afara intervalului se spune că este vorba de o extrapolare..



Acest gen de utilizare este frecvent întâlnită în probleme de etalonare.

Se măsoară cele două variabile pentru o serie de valori cunoscute, apoi pe baza dreptei de regresie se prezic valorile necunoscute.

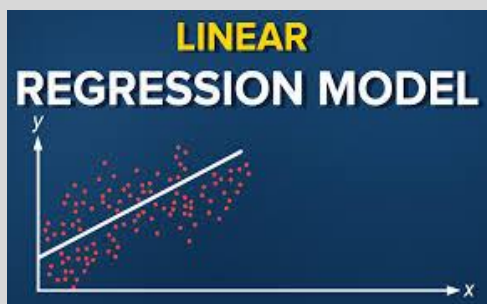


Cu atât mai mare este exactitatea prezicerii

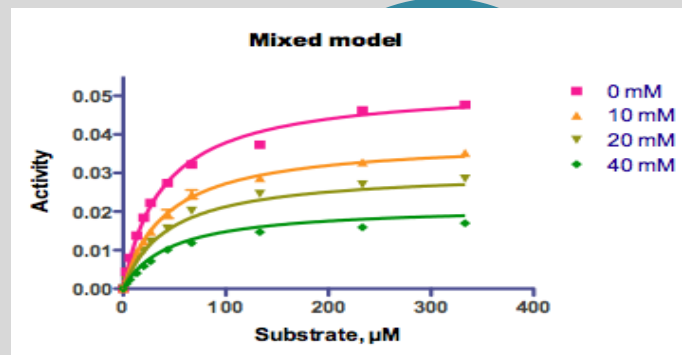
$Y = a + bx$  cu cât X ales este mai aproape de valoarea medie X .

La distanță mare de medie predicția utilizând modelul furnizat de regresia liniară devine riscantă.

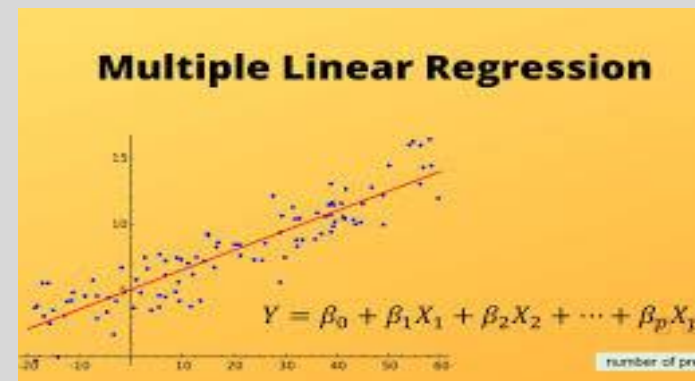
# Există 3 mari tipuri de modele de regresie



**REGRESIA  
LINIARĂ**



**REGRESIA  
NELINIARĂ**



**REGRESIA  
LINIARĂ  
MULTIPLĂ**

Relația dintre mai mult de  
două variabile cantitative

Continue

Binare

Predicția uneia  
în funcție de  
cealalte

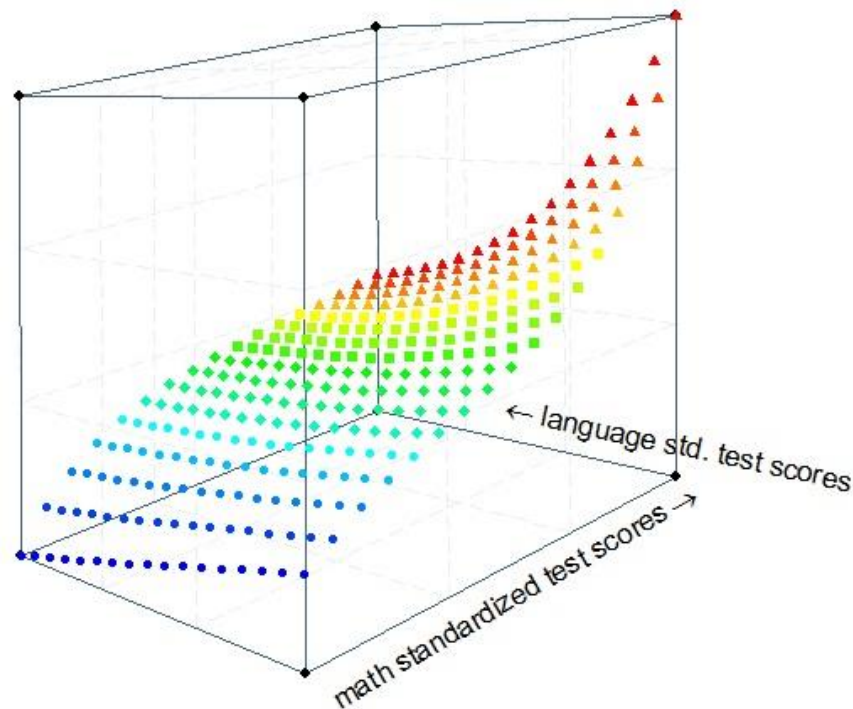
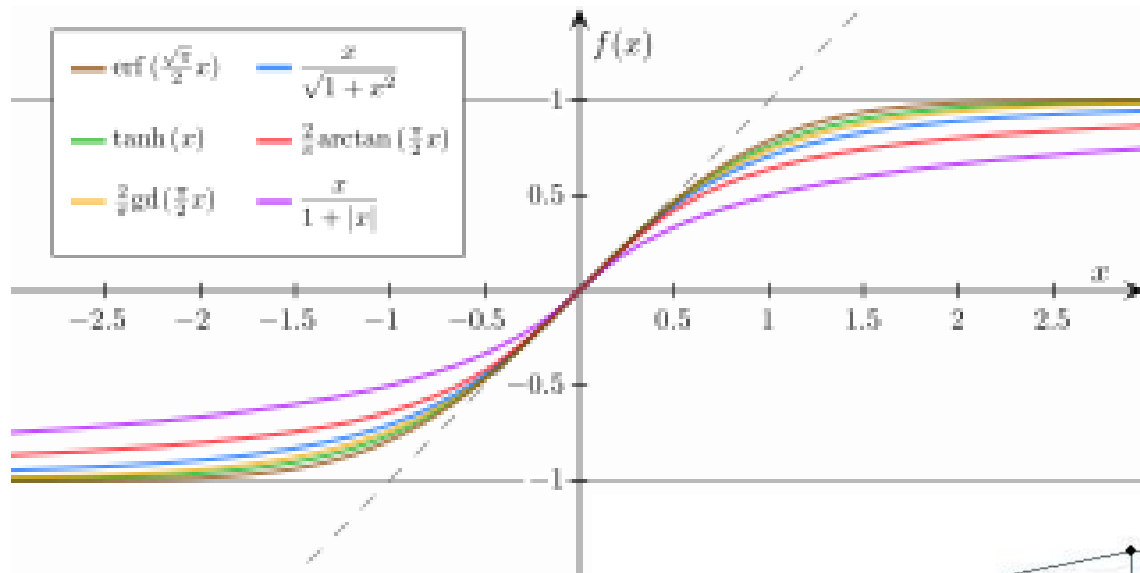
Regresii multiple

Regresii logistice  
multiple

# REGRESIA LINIARĂ MULTIPLĂ

Relația dintre mai mult de  
două variabile cantitative

## Exemple de regresii: sigmoidală (sus) și tridimensională (jos)



## REGRESII NELINIARE

Regresii polinomiale

Regresii gaussiene

Regresii sigmoidale


Regresii hiperbolice

Regresii exponentiale

Regresii logaritmice

Regresii tridimensionale

# REGRESIA LOGISTICĂ



În multe studii medicale se dorește obținerea de predicții pentru o variabilă dependentă discretă de tip binar (afectat da/nu, vindecat da/nu). Se urmărește care sunt variabilele care influențează variabila dependentă, în ce măsură contribuie acestea la predicție. Metoda matematică de regresie care aproximează o variabilă dicotomială prin una sau mai multe variabile cantitative sau calitative predictive poartă numele de regresie logistică. Pentru calculele necesare valorile posibile ale variabilei prezise vor fi notate cu 0 și cu 1.

**Interpretarea** rezultatelor acestui tip de regresie este mai puțin intuitivă, de exemplu panta dreptei „b” nu mai este interpretabilă ca rata schimbării lui Y în funcție de X, ci mai degrabă o rată a schimbării în funcție de logaritmul rației (raportului) șansei (ODDS RATIO) de modificare a lui X. Rația șansei (ODDS RATIO) se definește în acest caz ca raportul dintre șansa de a apărea evenimentul împărțită la șansa de a apărea evenimentul contrar sau ca raportul dintre șansa ca  $Y=1$  când  $X=1$  împărțită la șansa ca  $Y=1$  când  $X=0$  în cazul în care X este o variabilă calitativă. Validarea legăturii dintre predictor și variabila dependentă se face cu teste specifice de semnificație. Calculul coeficienților de regresie se face prin metoda iterativă a estimării verosimilității maxime (maximum likelihood estimation - MLE).



# ECUAȚIA DE REGRESIE

- Așa cum am văzut, coeficientul de corelație descrie intensitatea (tăria) asocierii între două variabile. Astfel, dacă două variabile sunt corelate, aceasta înseamnă că o modificare de o anumită mărime a valorii variabile independente va determina o modificare și în valoarea înregistrată la măsurarea celeilalte variabile.
- Pentru exemplul de mai sus, putem spune că o valoare mai mare a înălțimii copiilor este asociată cu o creștere, de o anumită factură, a spațiului mort anatomic.
- Dacă notăm cu Y variabila dependentă și cu X variabila independentă, putem afirma în consecință că relația poate fi descrisă ca o **regresie a lui Y în funcție de X**.
- Această relație poate fi reprezentată de o ecuație numită **ecuație de regresie**.
- În acest context termenul de **regresie** semnifică faptul că o anumită valoare a variabilei Y este o “funcție” de X, cu alte cuvinte se modifică odată cu modificarea valorii lui X, conform unei anumite ecuații mai mult sau mai puțin complexe.



# ECUAȚIA DE REGRESIE - continuare

- Cea mai simplă astfel de ecuație este ecuația dreptei (  $y = \beta x + \alpha$  ), iar regresia care folosește această ecuație poartă numele de **regresie liniară**.
- **Ecuația de regresie** ne arată cât de mult se schimbă valoarea variabilei Y în raport cu o anumită schimbare a variabile X și poate fi folosită pentru a trasa o așa-numită **linie de regresie**, în interiorul unei diagrame scatter-plot, iar cel mai simplu caz este cazul în care această linie este o **linie dreaptă**, caz în care se folosește termenul de **regresie liniară**.
- Direcția de “înclinare” a acestei linii de regresie depinde de faptul că avem de-a face cu o corelație pozitivă sau negativă. Astfel dacă cele două seturi de observații (x și y) cresc împreună (corelație pozitivă), linia de regresie va fi ascendentă de la stânga spre dreapta. Dacă valorile variabilei X cresc, iar valorile corespunzătoare ale variabilei Y descresc, înclinarea liniei de regresie va fi descendentă de la stânga spre dreapta.

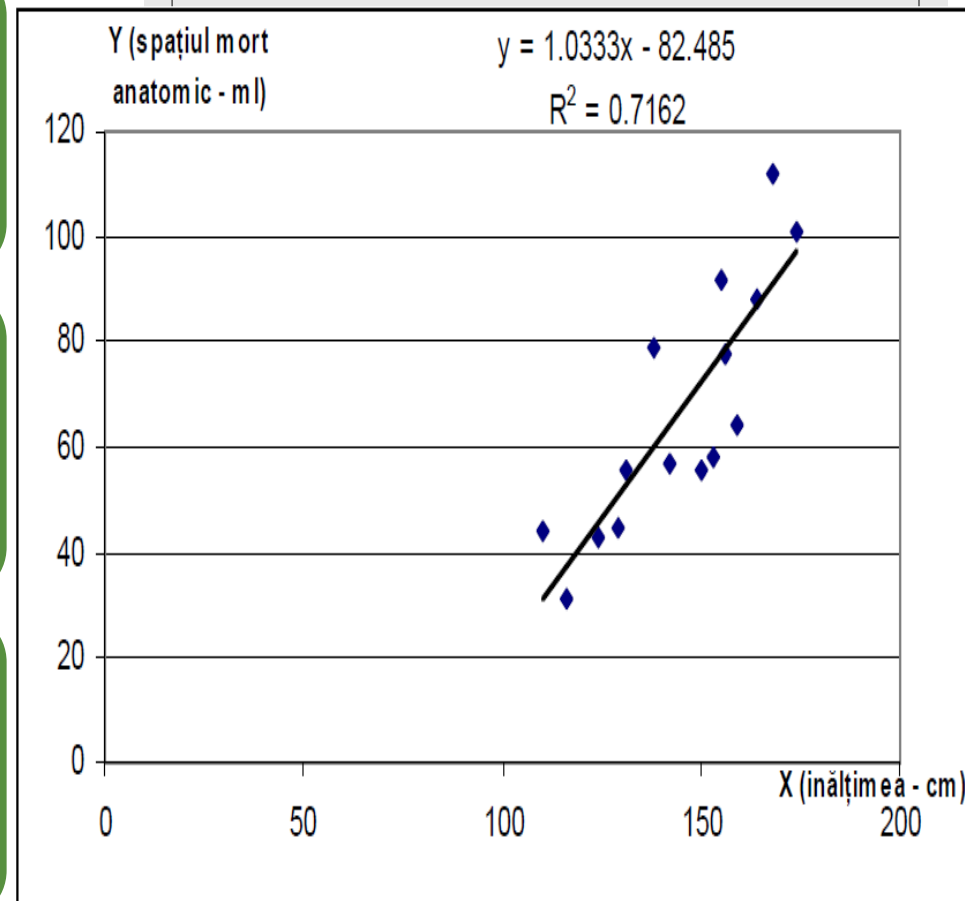
Din nefericire, în cazul regresiei liniare, de vreme ce avem de-a face cu o *dreaptă de regresie*, este foarte probabil ca ea să treacă prin relativ puține puncte reprezentate de noi în diagramă.

Fie ecuația dreptei de forma:  $y = \beta x + a$ . În momentul în care cunoaștem ecuația de regresie, pentru a putea trasa corect dreapta de regresie trebuie să ținem cont de cei doi coeficienți ai ecuației dreptei pentru a o putea trasa corect.

Primul este *interceptul*, adică punctul în care dreapta de regresie va intersecta axa OY și este dat de valoarea lui  $a$ . Pentru exemplul de mai sus, dreapta de regresie ar intercepta axa OY în dreptul valorii  $Y = -82,485$ .

Cel de-al doilea coeficient este  $\beta$ , și poartă numele de *pantă* a dreptei de regresie. Acest ultim parametru mai poartă numele și de coeficient de regresie și poate fi asimilat ca fiind mărimea modificării înregistrate în cazul valorii variabilei Y în urma modificării cu **o unitate** a valorii variabilei X ).

## ECUAȚIA DE REGRESIE - continuare



- Semnul pantei ecuației de regresie liniară ne arată clar dacă avem de-a face cu o corelație pozitivă sau negativă între cele două variabile,  $X$  și  $Y$ .
- În prezent programele de analiză statistică trasează automat dreapta de regresie, furnizând totodată și ecuația dreptei de regresie, respectiv coeficientul de determinare.
- În trecut, trasarea corectă a dreptei de regresie se făcea cu ajutorul metodei celor mai mici pătrate - least squares estimate (dreapta se trasa astfel încât suma pătratelor distanțelor de la punctele reprezentate în diagramă la dreapta de regresie să fie minimă).

Ținând cont de cele afirmate mai sus, **coeficientul de determinare  $r^2$**  este extrem de util deoarece:

— este o măsură a procentului variației ce poate fi “explicată” din totalul variației observate

↷ este o măsură a procentului în care varianța (fluctuația) unei Variabile (dependente) poate fi estimată (prezisă) din evoluția unei alte variabile (variabila independentă)

↻ este o măsură ce ne permite să determinăm cât de siguri putem fi în momentul în care facem “predicții” pentru un anumit model sau pentru date reprezentate într-o diagramă de tip scatter-plot

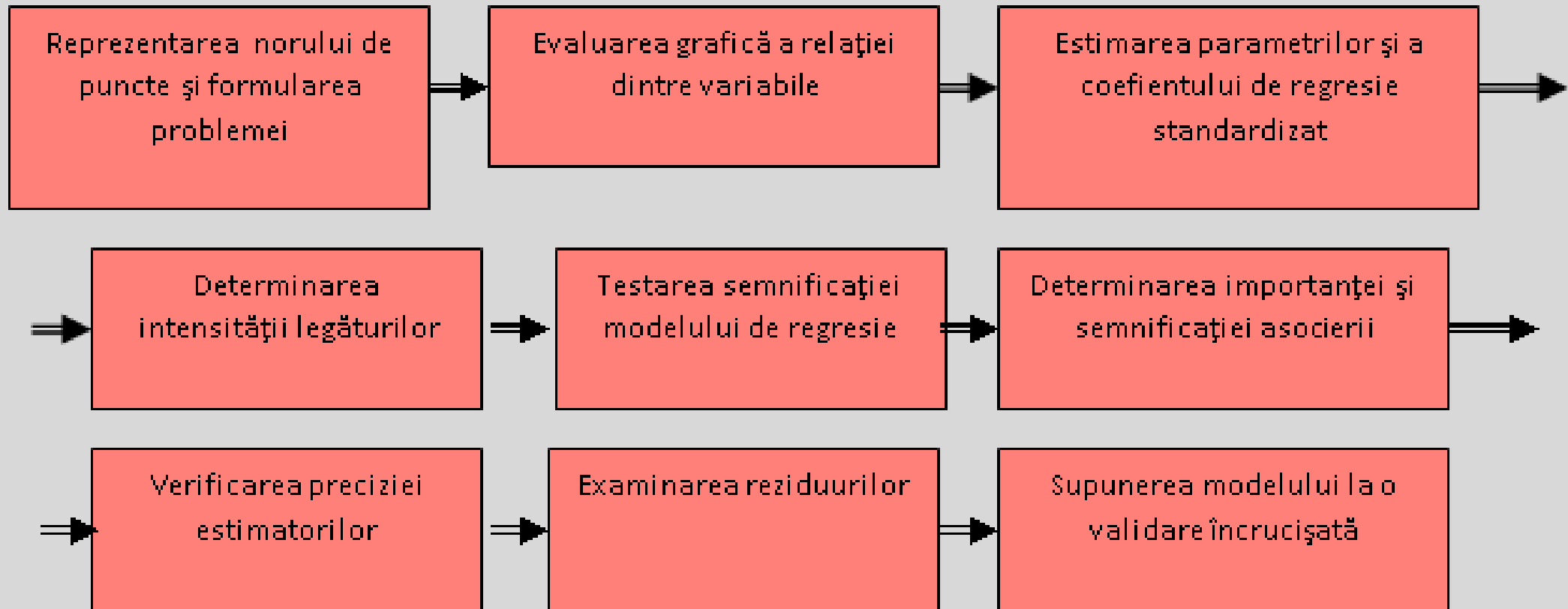
Ținând cont de cele afirmate mai sus, **coeficientul de determinare  $r^2$**  este extrem de util deoarece:

4 coeficientul de determinare, ce poate lua valori cuprinse între 0 și 1 ( $0 < r^2 < 1$ ) ne dă, în cazul regresiei liniare, o măsură a asocierii liniare dintre variabilele X și Y

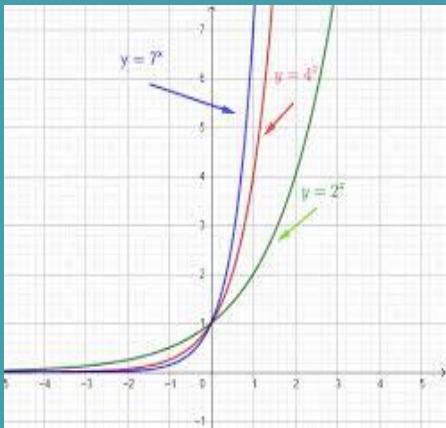
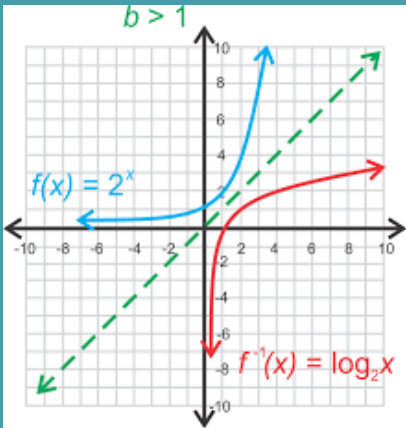
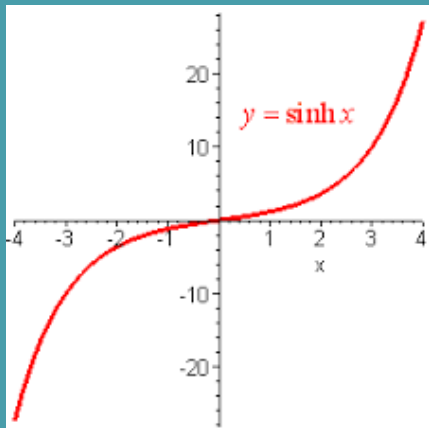
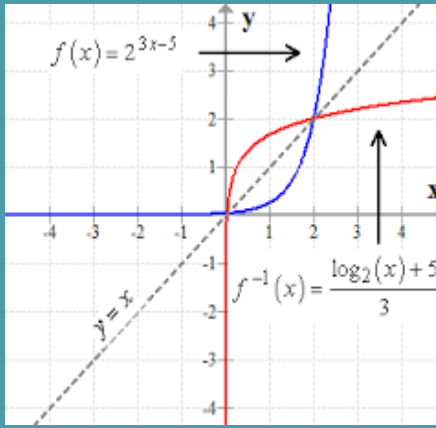
5 coeficientul de determinare reprezintă procentul de date care este cel mai apropiat de dreapta de regresie. De exemplu dacă avem un coeficient de corelație  $r = 0,922$  din care rezultă un coeficient de determinare  $r^2 = 0,850$ , aceasta înseamnă că 85% din totalul variației lui Y poate fi explicat printr-o relație liniară între X și Y, relație descrisă de ecuația de regresie. Restul de 15% din variație va rămâne neexplicată.

6 Coeficientul de determinare este, de asemenea, o măsură a gradului de exactitate (fidelitate) cu care o anumită linie de regresie reprezintă datele studiate. Astfel, dacă linia de regresie trece prin absolut toate punctele reprezentate în diagrama scatter, coeficientul de determinare va fi 1 și va putea explica întreaga variație. Cu cât linia de regresie este mai "îndepărtată" de puncte, cu atât coeficientul de va fi mai mic și un procent mai mare al variației nu va putea fi explicată.

# Etape parcurse în analiza de regresie:



# Transformarea proceselor neliniare în procese liniare

Tipuri de modele	Modelul exponențial	Modelul logaritmic	Modelul hiperbolic sau parabolic	Modelul exponențial inversat
Funcția inițială (neliniară)	$y_i = e^{ax_i+b}$	$y_i = a \ln x_i + b$	$y_i = x_i^a * b$	$y_i = e^{\frac{-a}{x_i}+b}$
Funcția liniarizată corespondentă	$\ln y_i = ax_i + b$	$\ln y_i = a \ln x_i + b$	$\ln y_i = a \ln x_i + \ln b$	$\ln y_i = \frac{-a}{x_i} + b$
Reprezentarea funcției inițiale				
Exemple de fenomene	Fenomene de creștere de tip geometric	Fenomene de elasticitate invers proporțională	Fenomene de elasticitate constantă	Fenomene de creștere de tip logistic

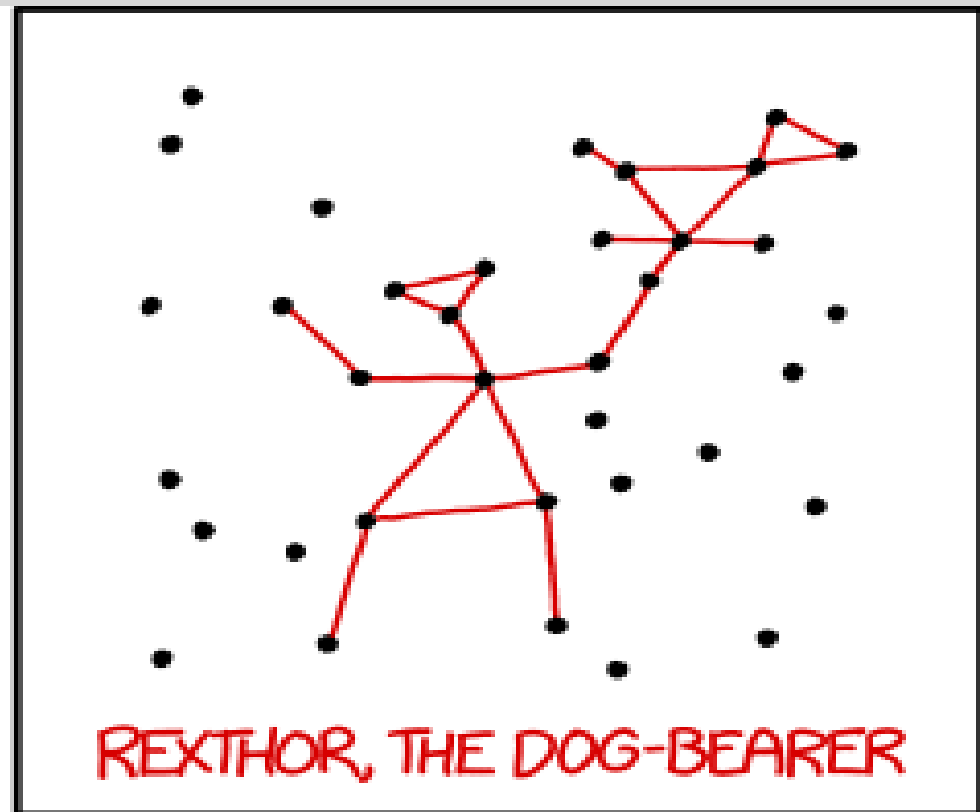
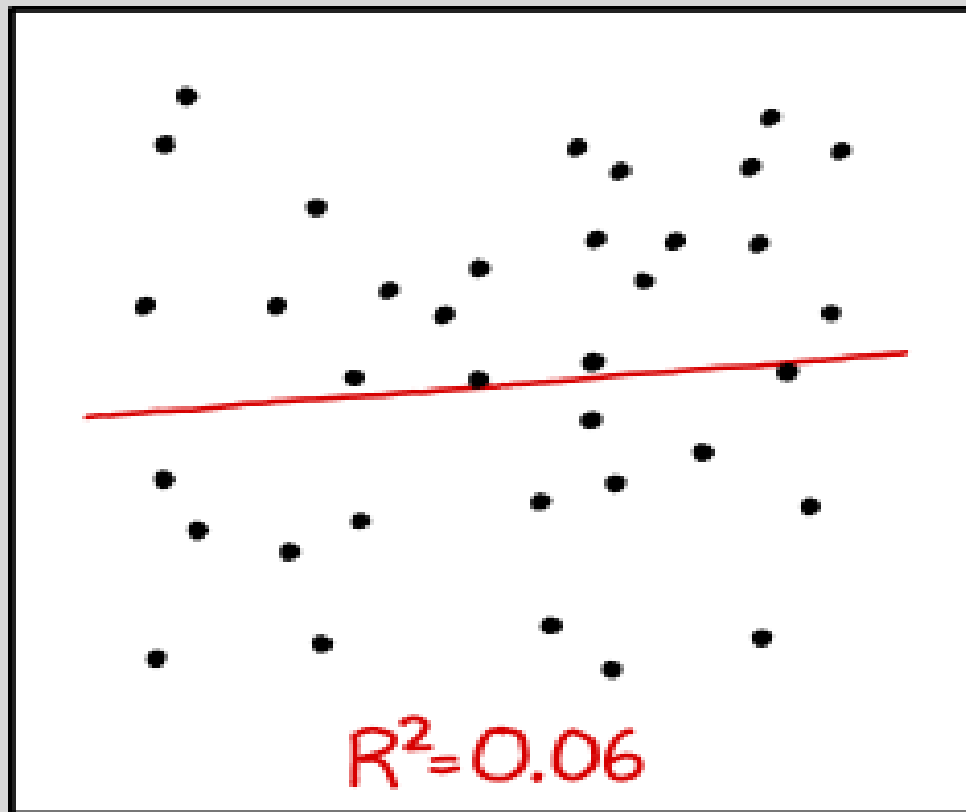


# Indicatori și noțiuni statistice asociate analizei de regresie

Indicatorul/ noțiunea utilizată	Descriere
<b>Model de regresie</b>	Ecuția modelului (pentru regresia bivariată) este: $Y_i = \beta_0 + \beta_1 X + e_i$ , unde Y = variabila dependentă, X = variabila independentă, $\beta_0$ = constanta ecuației de regresie (interceptul), $\beta_1$ – parametrul variabilei X (panta), $e_i$ – eroarea aferentă observației $i$ (abaterea valorii calculate a lui $Y_i$ de la valoarea observată)
<b>Coeficientul de determinatie</b>	Calculat ca pătratul coeficientului de corelație dintre X și Y ( $R^2$ ), măsoară intensitatea legăturii dintre cele două variabile și semnifică proporția variației variabilei dependente Y pe seama variabilei independente X (cazul regresiei bivariate).
<b>Coeficientul de determinatie multiplă</b>	Spre deosebire de coeficientul de determinatie calculat în cazul regresiei bivariate, acest coeficient ia în calcul toate variabilele independente incluse în modelul de regresie multiplă.
<b>Valoarea calculată (estimată sau previzionată)</b>	Reprezintă valoarea variabilei dependente obținută prin calcul folosind funcția de regresie $Y_i = b_0 + b_1 X$ , unde $b_0$ și $b_1$ sunt calculați ca estimatori ai parametrilor $\beta_0$ și $\beta_1$ .

# Indicatori și noțiuni statistice asociate analizei de regresie

Indicatorul/ noțiunea utilizată	Descriere
<b>Nor de puncte</b>	Reprezentarea grafică a valorilor celor două variabile pentru toți indivizii sau observațiile din studiu.
<b>Valori reziduale</b>	Calculate ca diferență între valorile observate și valorile calculate ale variabilei dependente.
<b>Coeficient de regresie</b>	Este coeficientul <b><math>b_1</math></b> al funcției de regresie (pentru regresia bivariată) când variabilele din model nu au fost standardizate și arată cu cât se modifică variabila dependentă la o schimbare cu o unitate a variabilei independente. Pentru datele standardizate, poartă denumirea de <b>coeficient de regresie standardizat</b> sau <b>coeficient beta</b> .
<b>Coeficient de regresie parțiali</b>	Reprezintă coeficienții variabilelor independente din ecuația de regresie și arată impactul fiecăreia asupra variabilei dependente.
<b>Eroarea standard</b>	Este abaterea standard a coeficientului de regresie $b_1$ .
<b>Suma pătratelor abaterilor</b>	Este o măsură a erorii totale și reprezintă distanțele tuturor punctelor (valorilor observate) la dreapta de regresie (valori calculate) ridicate la pătrat și însumate.
<b>Teste statistice</b>	Testul t, testul F, testul Durbin – Watson (sunt regăsite detaliate în capitolul IV, paragraful 4.1.)



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER  
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE  
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.