# ECONOMETRIE

## CURS 2

- note de curs -

IAŞI
- 2023-

## C2. REGRESIA LINIARĂ SIMPLĂ (1)

**TEMATICA CURS 02**

Prezentarea modelului - regresia empirică și cea teoretică

Estimarea punctuală şi prin interval de încredere a parametrilor - MCMMP

Testarea parametrilor

Probleme specifice utilizând SPSS si Excel

## 1.Scurt istoric al regresiei

**1886** - *Francis Galton,*. “Family Likeness in Stature”, *Proceedings of Royal Society*, London, vol. 40, 1886, pp. 42-72

**1897** - *G. U. Yule*, "On the Theory of Correlation", Journal of the Royal Statistical Society , pp. 812-54.

**1903** - *Karl Pearson, G. U. Yule, Norman Blanchard, and Alice Lee,* "The Law of Ancestral Heredity", *Biometrika*

**1925** - *R.A. Fisher, Statistical Methods for Research Workers*

## 2. Noţiuni (1)

**Regresia** este o *legătură statistică* între două sau mai multe variabile statistice.

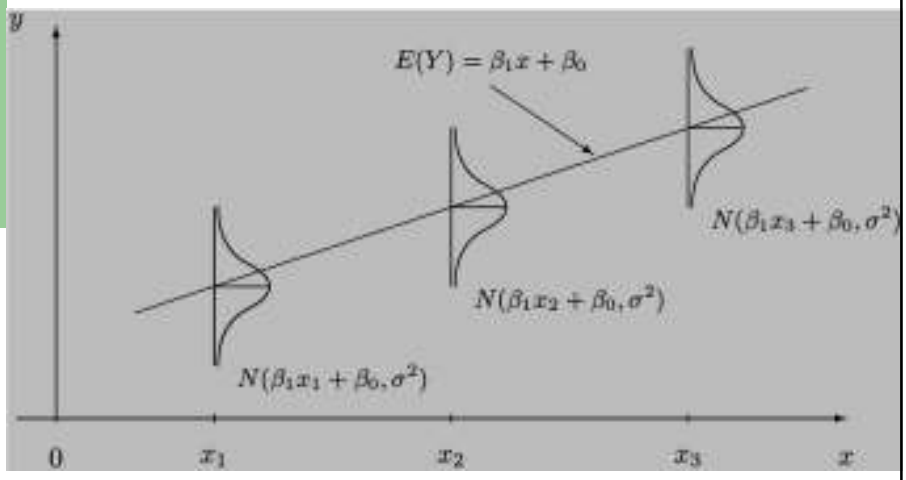
În descrierea *legăturilor statistice pentru variabilele dependente* utilizăm *variabile aleatoare (stohastice),* ceea ce înseamnă că acestora le corespund *distribuţii de probabilitate* (fig. 1).

$x_i \rightarrow Y: \begin{pmatrix} y_1 \dots y_j \dots y_m \\ n_i \dots n_j \dots n_m \end{pmatrix}$ **=> M(Y | $x_i$) = f($x_i$) => $y_i$=f($x_i$)+$\varepsilon_i$**

X – *variabilă independentă* - variabilă nestohastică

Y – *variabilă dependentă* – variabilă stohastică (prezintă distribuţii pentru fiecare valoare a lui X)

## 2. Noţiuni (2) - Fig.1




În *legăturile de tip funcţional* unei valori ***i*** se asociază o altă valoare şi nu o distribuţie de probabilitate.

**$x_i \rightarrow y_i$ => $f(x_i)=y_i$**

**Analiza de regresie** *studiază forma legăturii dintre una sau mai multe variabile* => **Model de regresie**

**Analiza de corelaţie** *studiază intensitatea legăturii* dintre una sau mai multe variabile puse în relație printr-un *model de regresie*.

# Modele de regresie

| Modele de regresie | Simple | Multiple |
|---|---|---|
| Liniare | $C = \beta_0 + \beta_1 X + \varepsilon$ | $C = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ |
| Neliniare | $C = e^{\beta_0+\beta_1 X+\varepsilon}$ | $C = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\varepsilon}$ |

# 3. Modelul de regresie liniară simplă

$$\varepsilon_i = y_i - M(Y \mid X=x_i) \Leftrightarrow y_i = M(Y \mid X=x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

> $M(Y \mid X=x_i) = \beta_0 + \beta_1 x_i$ – *media condiționată de $X=x_i$, a variabilei stohastice Y,*

> $\beta_0 = f(0)$ – parametrul "*intersecţia dreptei de regresie liniară cu axa OY*" (*engl.* intercept)

> $\beta_1$ – parametrul "*panta a dreptei*" care reprezintă variaţia absolută ***in medie*** a variabilei Y atunci când variabila X creşte cu o unitate ($\beta_1 = \Delta Y / \Delta X$):

- $\beta_1 > 0$: *legătură directă* între variabile, Y variază în acelaşi sens cu X
- $\beta_1 < 0$: *legătură inversă* între variabile, Y nu variază în acelaşi sens cu X

## 3. Modalități de scriere ale modelului liniar simplu

Dacă este scris pentru *valorile variabilelor*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Dacă este scris pentru *variabile în general*

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

## 4. Componentele modelului de regresie

**A. Componenta deterministă ($\beta_0 + \beta_1 x_i$)**

**B. Componenta aleatoare ($\varepsilon_i$).**

**Factorii care influențează componenta aleatoare**:

- *natura fenomenului studiat;*
- *specificarea modelului ;*
- *erorile de măsurare s.a.*

## 5. Ipoteze clasice ale modelului de regresie

Ipotezele modelului de regresie vizează *variabila rezidual*ă şi *variabila independentă*.

Cele mai importante **ipoteze cu privire la variabila rezidual**ă sunt:

- *normalitatea erorilor* :$\varepsilon_i \sim N(\mu_i, \sigma_i^2)$, adică variabila reziduală urmează o lege de repartiţie normală de medie $\mu_i$ şi varianţă $\sigma_i^2$;
- *media erorilor de modelare este nulă:* $M(\varepsilon_i)=0 \Rightarrow \varepsilon_i \sim N(0, \sigma_i^2)$
- *homoscedasticitate*:$V(\varepsilon_i) = M(\varepsilon_i^2) = \sigma^2$,adică varianţa erorii este constantă la nivelul distribuţiilor condiţionate de tipul $Y_i | X = x_i$ -> $\varepsilon_i \sim N(0, \sigma^2)$
- *necorelarea erorilor:* $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j;\ i, j = \overline{1,n}$ ,adică erorile nu se influenţează reciproc;
- *lipsa corelaţiei dintre variabila independentă şi variabila eroare*: $cov(\varepsilon_i, x_i) = 0$.

## 6. Exemple de modele liniare simple

**Funcţia de consum**

- *cererea sau consumul populaţiei* în funcţie de *venit*:

$C_i = \beta_0 + \beta_1 V_i + \varepsilon_i$

unde parametrul $\beta_1$ arată cu cât creşte **in medie consumul** unui anumit produs ($C_i$) la o creştere cu o unitate a venitului şi este de regulă pozitiv.

**Legea cererii**

- *cererea populaţiei* pentru o gamă de produse în funcţie de *preţul* acestora:

$C_i = \beta_0 + \beta_1 P_i + \varepsilon_i$, unde parametrul $\beta_1$ este de regulă negativ şi arată cu cât scade **in medie cererea** la o creştere a preţului cu o unitate.

**STOP**

## 7. Estimarea punctuală parametrilor modelului de regresie prin MCMMP

**MCMMP** ***(engl. Method of Ordinary Least Sqares - OLS)***

Fie: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ - *valorile estimate* (teoretice ale lui Y)

și $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$ - *valorile reale, înregistrate*

Erorile estimate pot fi obtinute ca diferenta intre **valorile reale** si valorile **teoretice**:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Pentru a ușura notațiile în procesul de estimare vom utiliza direct **notațiile pentru estimații**. Astfel relația de mai sus se va scrie:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 - \min.$$

Notăm $S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$

Rezolvarea acestei probleme de minim presupune îndeplinirea a *două condiţii*:

1. *Anularea derivatelor parţiale de ordinul I ale lui S* în raport $\mathbf{b_0}$ şi $\mathbf{b_1}$:

$$\begin{cases} \dfrac{\partial S}{\partial b_0} = 2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(-1) = 0 \\ \dfrac{\partial S}{\partial b_1} = 2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(-x_i) = 0 \end{cases} \Rightarrow \begin{cases} n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases}$$

$$b_0 = \frac{\Delta b_0}{\Delta} = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

sau $b_0 = \bar{y} - b_1 \bar{x}$

$$b_1 = \frac{\Delta b_1}{\Delta} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

2. *Matricea derivatelor parţiale de ordinul doi să fie pozitiv definită*:

$$\det\begin{pmatrix} \frac{\partial^2 S}{\partial^2 b_0} = 2n & \frac{\partial^2 S}{\partial b_0 \partial b_1} = 2\sum x_i \\ \frac{\partial^2 S}{\partial b_0 \partial b_1} = 2\sum x_i & \frac{\partial^2 S}{\partial^2 b_1} = 2\sum x_i^2 \end{pmatrix} > 0 <=> \det\begin{pmatrix} 2n & 2\sum x_i \\ 2\sum x_i & 2\sum x_i^2 \end{pmatrix} > 0$$

$$<=> \det\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} > 0$$

Matricea derivatelor partiale de ordin doi este pozitiv definită deoarece $n\sum x_i^2 - \left(\sum x_i\right)^2 = n^2\sigma^2 > 0$

## 7. Proprietăţile estimatorilor parametrilor modelului de regresie

*Estimatorii* parametrilor modelului de regresie sunt variabile de selecţie care:

- urmează o *distribuţie normală*: $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right)$, $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$

- sunt *nedeplasaţi*: $M\left(\hat{\beta}_0\right) = \beta_0$, $M\left(\hat{\beta}_1\right) = \beta_1$

- *convergenţi* (in probabilitate): $\left(\hat{\beta}_0\right)_{n\in N} \xrightarrow{p} \beta_0$, $\left(\hat{\beta}_1\right)_{n\in N} \xrightarrow{p} \beta_1$

- *eficienţi*: dintre toţi estimatorii posibili pentru $\beta_1$, $\hat{\beta}_1$ are varianţa cea mai mică

## 8. Estimarea prin interval de încredere a parametrilor modelului de regresie liniară

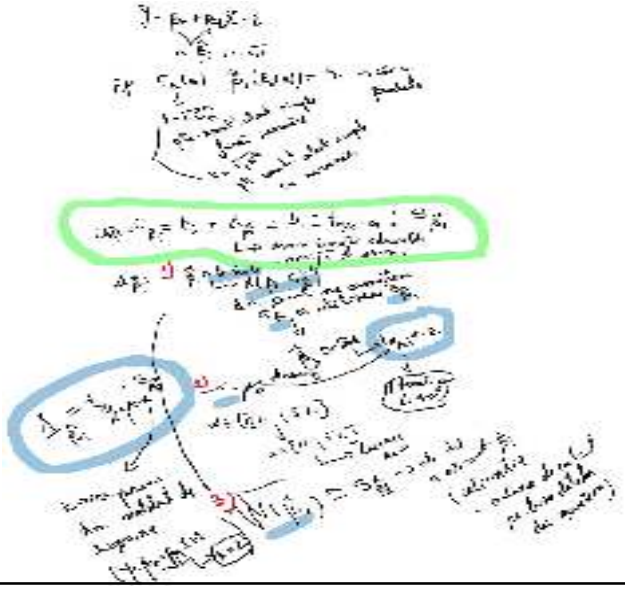
Atât pentru $\beta_0$, cât şi pentru $\beta_1$, *intervalele de încredere pentru* se vor construi astfel:

$$\beta_0 \in \left[ b_0 - t_{\alpha/2,n-k} s_{\hat{\beta}_0} , b_0 + t_{\alpha/2,n-k} s_{\hat{\beta}_0} \right]$$

$$\beta_1 \in \left[ b_1 - t_{\alpha/2,n-k} s_{\hat{\beta}_1} , b_1 + t_{\alpha/2,n-k} s_{\hat{\beta}_1} \right]$$

unde **k** = *numărul parametrilor estimaţi din model* (pentru modelul liniar k=2)

## Precizari suplimentare



**Vezi exemplu excel (n=19 iar α=5%): Y-> greutatea (kg) si X-> inaltimea (cm) Y=β0+β1X+ε -> K=2**

| | Coefficients ($b_i$) | Standard Error ($s_{\hat{\beta}_i}, i=0,1$) | t Stat ($t_{calc}$) | P-value (sig.) | Lower 95% ($L_i$) | Upper 95% (Ls) |
|---|---|---|---|---|---|---|
| *Intercept* **(β0)** | **-151.816** ($b_0$) | 61.808($s_{\hat{\beta}_0}$) | -2.456 | 0.025 | -282.219 | -21.412 |
| *Inaltimea (X)* **(β1)** | **1.266** ($b_1$) | 0.364($s_{\hat{\beta}_1}$) | 3.483 | 0.003 | 0.499 | 2.034 |

$t_{\alpha/2,n-k}$ -> **aceasta este o valoare teoretica si se ia din tabelul valorilor statisticii Student si nu din tabelul de mai sus!!!!**

$\beta_0 \in \left[b_0 - t_{\alpha/2,n-k} s_{\hat{\beta}_0}, b_0 + t_{\alpha/2,n-k} s_{\hat{\beta}_0}\right]$ = -151.816± 2,11*61,808 -> vezi $L_i$ si $L_s$ din tabelul de mai sus

$\beta_1 \in \left[b_1 - t_{\alpha/2,n-k} s_{\hat{\beta}_1}, b_1 + t_{\alpha/2,n-k} s_{\hat{\beta}_1}\right]$ = 1,266± 2,11*0,364 -> vezi $L_i$ si $L_s$ din tabelul de mai sus

$t_{\alpha/2,n-k} = t_{0,025;(19-2)} = t_{0,025;17} = 2,11$

Student t Distribution Table (thoughtco.com)

# 5. Testarea parametrilor modelului liniar

## 1. Formularea ipotezelor

**pentru $\beta_0$**
**$H_0$:** $\beta_0=0$
**$H_1$:** $\beta_0 \neq 0$

**pentru $\beta_1$**
**$H_0$:** $\beta_1=0$
**$H_1$:** $\beta_1 \neq 0$

## 2. Fixarea pragului de semnificaţie

**α=0,05**

## 3. Alegerea si calcul statisticii test

$t_{\beta_0} = \frac{b_0}{s_{\hat{\beta}_0}} ? t_{teoretic} = t_{\alpha/2,n-2}$ $\qquad$ $t_{\beta_1} = \frac{b_1}{s_{\hat{\beta}_1}} ? t_{teoretic} = t_{\alpha/2,\ n-2}$

## 4. Criterii de decizie:

**$|t_{calc}| \leq t_{teoretic} = t_{\alpha/2,\ n-2}$** sau **sig.≥α**=> A$H_0$ cu o probabab. de 1-α.

**$|t_{calc}| > t_{teoretic} = t_{\alpha/2,\ n-2}$** sau **sig.<α** => R$H_0$ cu un risc asumat α.

| | Coefficients ($b_i$) | Standard Error ($s_{\hat{\beta}_i}, i=\overline{0,1}$) | t Stat ($t_{calc}$) | P-value (sig.) | Lower 95% ($L_i$) | Upper 95% (Ls) |
|---|---|---|---|---|---|---|
| *Intercept* **(β0)** | **-151.816** | 61.808 | -2.456 | 0.025 | -282.219 | -21.412 |
| *Inaltimea (X)* **(β1)** | **1.266** | 0.364 | 3.483 | 0.003 | 0.499 | 2.034 |

**1. Formularea ipotezelor**

**pentru $\beta_0$**
**$H_0$:** $\beta_0=0$
**$H_1$:** $\beta_0 \neq 0$

**pentru $\beta_1$**
**$H_0$:** $\beta_1=0$
**$H_1$:** $\beta_1 \neq 0$

**2. Fixarea pragului de semnificaţie**
**α=0,05**

**3. Alegerea si calcul statisticii test**

$$t_{calc\, b_0} = \frac{b_0}{s_{\hat{\beta}_0}} \sim t_{\alpha/2,\, n-2}$$

$$t_{calc\, b_1} = \frac{b_1}{s_{\hat{\beta}_1}} \sim t_{\alpha/2,\ n-2}$$

**4. Criterii de decizie**:

**$|t_{calc\ \beta 0}| = |-2,456| > t_{teoretic} = t_{0,025;\ 17} = 2,11$**

**sig.=0.025<α=0,05**

=> $RH_0$ cu un risc asumat α

**$|t_{calc\ \beta 1}| = |-2,456| \leq t_{teoretic} = t_{0,025;\ 17} = 2,11$**

**sig.=0.003<α=0,05**

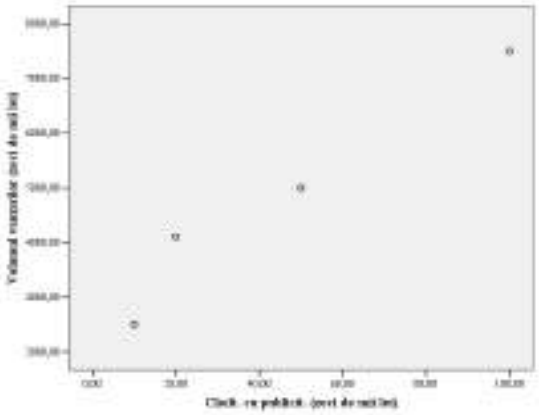
=> $RH_0$ cu un risc asumat α .

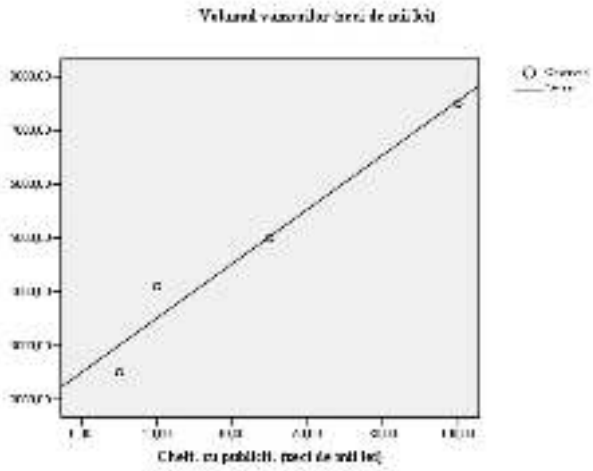
## 5. Probleme specifice analizei de corelaţie şi regresie

**Analiza de regresie si corelaţie în SPSS**

Se consideră datele cu privire la *Valoarea vânzărilor* şi *Cheltuielile cu publicitatea* pentru un eşantion de 4 firme.
Datele sunt prezentate în tabelul alăturat.

| $x_i$ | $y_i$ |
|---|---|
| 10 | 2500 |
| 20 | 4100 |
| 50 | 5000 |
| 100 | 7500 |
| **180** | **19100** |





**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .977[a] | .954 | .931 | 550.55630 |

a. Predictors: (Constant), X

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 12501275.51 | 1 | 12501275.51 | 41.243 | .023[b] |
| | Residual | 606224.490 | 2 | 303112.245 | | |
| | Total | 13107500.00 | 3 | | | |

a. Dependent Variable: Y

b. Predictors: (Constant), X

**Correlations**

| | | X | Y |
|---|---|---|---|
| X | Pearson Correlation | 1 | .977* |
| | Sig. (2-tailed) | | .023 |
| | N | 4 | 4 |
| Y | Pearson Correlation | .977* | 1 |
| | Sig. (2-tailed) | .023 | |
| | N | 4 | 4 |

*. Correlation is significant at the 0.05 level (2-tailed).

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 2502.041 | 449.379 | | 5.568 | .031 | 572.821 | 4431.260 |
| | X | 50.510 | 7.865 | .977 | 6.422 | .023 | 16.689 | 84.351 |

a. Dependent Variable: Y