**Statistică multivariată**

# Lucrarea nr. 8 — Regresia liniară multiplă
# - Excel, SPSS -

## *A. Noțiuni teoretice*

Regresia liniară, prin metoda celor mai mici pătrate, este metoda de modelare cea mai des utilizată. Este metoda denumită "regresie", "regresie liniară", "regresie multiplă" sau "cele mai mici pătrate" atunci când se construieşte un model.

Scopul regresiei multiple (termen utilizat de Pearson, 1908) este de a evidenția relația dintre o variabilă dependentă (explicată, endogenă, rezultativă) şi o mulțime de variabile independente (explicative, factoriale, exogene, predictori). Prin utilizarea regresiei multiple se încearcă, adesea, obținerea răspunsului la una dintre întrebările: "care este cea mai bună predicție pentru …?", "cine este cel mai bun predictor pentru …?" .

De reținut că metoda regresiei multiple este generalizată prin teoria "modelului liniar general", în care se permit mai multe variabile dependente simultan şi, de asemenea, variabile factoriale care nu sunt independente liniar.

Clasa modelelor liniare poate fi exprimată prin

$$y = \mathbf{x}\,\boldsymbol{\alpha} + \varepsilon$$

unde

- y este variabila dependentă (explicată, endogenă, rezultativă),
- **x** este vectorul variabilelor independente (explicative, exogene), de dimensiune $1 \times p$,
- $\boldsymbol{\alpha}$ este vectorul coeficienților, de dimensiune $p \times 1$, parametrii modelului,
- $\varepsilon$ este o variabilă, interpretată ca eroare (perturbare, eroare de măsurare etc.).

Cu alte cuvinte,

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p + \varepsilon$$

care exprimă relația liniară dintre y şi x.

**Observații**. 1. Liniaritatea relației se referă la coeficienți şi nu la variabile. Astfel, modelul

$$y = \alpha_1 x_1^2 + \alpha_2 \sqrt{x_2} + \alpha_3 \frac{1}{x_3} + \varepsilon$$

este tot un model liniar.

2. Considerând că $x_1$ este constant egală cu 1, se obține un model liniar care include un termen constant (termenul liber al modelului).

3. Pentru $p = 2$ şi $x_1 \equiv 1$ se obține modelul liniar simplu, dreapta de regresie.

4. Utilitatea principală a unui model liniar este aceea a predicției valorii lui $y$ din valorile cunoscute ale variabilelor **x**.

Presupunem că avem un set de $n$ observații efectuate asupra variabilelor implicate în model. Prin urmare dispunem de $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, $i = 1, 2, \ldots, n$. Notând cu **y** vectorul de tip $n \times 1$ având drept componente valorile măsurate pentru variabila $y$, cu **X** matricea $(x_{ij})_{n \times p}$ a valorilor măsurate pentru variabilele **x** şi cu $\boldsymbol{\varepsilon}$ vectorul de tip $n \times 1$ având drept componente valorile erorilor, modelul se rescrie în relația matriceală:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

**Ipoteze inițiale.** În tot ceea ce urmează se presupun îndeplinite ipotezele:

1. Matricea de experiențe, $n$ observații pentru $p$ variabile, este fixată: $\mathbf{X}_{n\times p}$ nu este stohastică. În plus, $n >> p$.
2. **X** este de rang $p$ (coloanele sunt liniar independente – formează o bază a unui spațiu vectorial $p$-dimensional).
3. **a.** Vectorul de perturbații (n-dimensional) **ε** constă din n variabile aleatoare independente cu media 0 şi aceeaşi dispersie:

   $\text{Exp}(\boldsymbol{\varepsilon}) = 0$

   $\text{Var}(\boldsymbol{\varepsilon}) = \text{Exp}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}_n$, unde $\sigma^2$ este un parametru necunoscut,

   sau,

   **b.** Vectorul **ε** este o v.a. $n$-dimensională normală

   $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$.

De remarcat că ultima ipoteză, a normalității, este, mai degrabă, o ipoteză simplificatoare decât una restrictivă, cum sunt primele două. Aceasta deoarece erorile se datorează, în general, în procesele studiate, acțiunilor simultane ale unor factori aleatorii, ceea ce prin teorema de limită centrală conduce la concluzia că **ε**, ca sumă a lor, tinde spre o repartiție normală.

Problemele principale urmărite sunt:

- estimarea coeficienților $\alpha$,
- calitatea estimării,
- verificarea ipotezelor,
- calitatea predicției,
- alegerea modelului.

## Estimația prin cele mai mici pătrate

Numim **estimație** (ajustare) a modelului orice soluție {**a**, **e**} a sistemului

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e}.$$

Este de remarcat că sistemul conține $n$ ecuații şi $p + n$ necunoscute, deci admite o infinitate de soluții.

Numim **estimație prin cele mai mici pătrate**, acea soluție **a** care minimizează suma pătratelor erorilor $e_i$, adică

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a_1 x_{i1} + a_2 x_{i2} + \ldots + a_p x_{ip})]^2 .$$

Cum $\sum_{i=1}^{2} e_i^2 = \mathbf{e}'\mathbf{e}$ este o funcție de coeficienții $a$, o condiție necesară pentru atingerea maximului este

$$\frac{\partial}{\partial \mathbf{a}} (\mathbf{e}'\mathbf{e}) = 0 .$$

Se obține

$$\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

şi se demonstrează că este îndeplinit criteriul de minim şi că este singura valoare cu această proprietate adică valorile determinate reprezintă estimația prin cele mai mici pătrate a coeficienților modelului liniar.

Ecuația

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_p x_p$$

se numeşte **ecuația de regresie multiplă**.

Înlocuind în această relație valori pentru variabilele independente $x_i$ se obține valoarea prognozată pentru variabila dependentă $y$.

**Interpretarea coeficienților**

Un coeficient $a_i$ are interpretarea: modificarea cu 1 a valorii variabilei $x_i$ produce o modificare a valorii $y$ cu $a_i$ unități. Deoarece scalele de măsură sunt, în general, diferite, interpretarea în acest sens a coeficienților poate deforma imaginea importanței variabilelor independente în model. Din acest motiv se introduc coeficienții de regresie standardizați definiți drept coeficienții de regresie estimați ai modelului:

$$\tilde{y} = \beta_1\tilde{x}_1 + \beta_2\tilde{x}_2 + \ldots + \beta_p\tilde{x}_p$$

în care nu există termen liber, iar variabilele $\tilde{y}$ şi $\tilde{x}_i$ sunt variabilele standardizate, prin *standardizare* înțelegându-se transformarea de tipul $\tilde{x} = \frac{x - \bar{x}}{s_x}$.

Coeficienții de regresie standardizați au interpretarea: modificarea cu o abatere standard a valorii variabilei $x$ produce o modificare cu $\beta_i$ abateri standard a valorii variabilei dependente. În acest fel, mărimea coeficienților standardizați reflectă importanța variabilelor independente în predicția lui $y$.

**Distribuția estimatorului**

$\text{Exp}(\mathbf{a}) = \alpha$
$\text{Var}(\mathbf{a}) = \sigma^2(\mathbf{X'X})^{-1}$.

**Estimația dispersiei erorilor ($\sigma^2$)**

Notând cu $\hat{y}$ valoarea ajustată, dată de ecuația de regresie, pentru o realizare a vectorului **x,** considerată la estimarea parametrilor, se obține eroarea de ajustare, notată cu $e$:

$$e_i = y_i - \hat{y}_i,\ i = 1,\ldots,n.$$

Erorile de ajustare sunt denumite uzual *reziduuri* şi analiza lor este o parte importantă studiului calitativ al ecuației de regresie. Este evident că reziduurile constituie estimații ale erorilor ε. Se demonstrează că

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p}$$

este o estimație nedeplasată a dispersiei necunoscute $\sigma^2$. Este de notat că numitorul este egal cu numărul gradelor de libertate a sumei de la numărător ($n$ observații din care am obținut $p$ estimații).

## Precizia ajustării

Reziduuri mici exprimă o ajustare mai bună a datelor experimentale, dar stabilirea unui criteriu care să indice cât de mici trebuie să fie reziduurile pentru ca regresia să fie acceptată este o problemă dificilă.

Pentru a obține o măsură a preciziei ajustării se pleacă de la identitatea

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

care, prin reorganizarea termenilor, produce

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

Se poate demonstra că are loc identitatea:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2.$$

Această relație arată că variația valorilor observate în jurul valorii medii se descompune într-un termen ce exprimă variația valorilor estimate în jurul mediei şi într-un termen datorat reziduurilor ajustării. Prin urmare, regresia estimată va fi cu atât mai bună cu cât ultimul termen va fi mai mic, sau cu cât variația valorilor estimate va fi mai apropiată de variația valorilor observate. Se alege drept indicator sintetic de precizie a ajustării raportul

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}.$$

Pentru o bună ajustare a ecuației de regresie la datele experimentale, trebuie ca acest raport să fie apropiat de 1.

Cantitatea $R^2$ se numeşte **coeficientul de determinare** şi, exprimat procentual, arată cât din varianța variabilei dependente este explicată de ecuația estimată. Este un indicator de asociere având atributul PRE,

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

deci poate fi interpretat şi în următorul sens: cu cât se îmbunătățeşte prognoza valorilor y prin considerarea modelului estimat.

Se arată că $R^2$ creşte prin includerea mai multor variabile în model, astfel încât are loc o supraestimare în cazul modelelor extinse. O soluție propusă este *ajustarea coeficientului de determinare* prin

$$\overline{R}^2 = R^2 - \frac{p-1}{n-p}(1-R^2).$$

**Coeficientul de corelație multiplă**

Ca măsură a asocierii dintre $y$ şi ansamblul variabilelor $x$ se introduce coeficientul de corelație multiplă, notat cu $R$. Poate fi definit drept coeficientul maxim de corelație simplă (Pearson) dintre $y$ şi o combinație liniară de variabile $x$. Astfel se explică faptul că valoarea calculată a lui $R$ este întotdeauna pozitivă şi tinde să crească o dată cu mărirea numărului de variabile independente.

Metoda celor mai mici pătrate poate fi astfel gândită ca o metodă care maximizează corelația dintre valorile observate şi valorile estimate (acestea reprezentând o combinație liniară de variabile $x$). O valoare $R$ apropiată de 0 denotă o regresie nesemnificativă, valorile prognozate de regresie nefiind mai bune decât cele obținute printr-o ghicire aleatorie (sau bazate doar pe distribuția lui $y$).

Deoarece $R$ tinde să supraestimeze asocierea dintre $y$ şi **x**, se preferă indicatorul definit anterior, coeficientul de determinare, $R^2$, care este pătratul coeficientului de corelație multiplă.

## Testarea ipotezelor

Notăm

$$SP_g = \sum_i (y_i - \bar{y})^2, \quad SP_{reg} = \sum_i (\hat{y}_i - \bar{y})^2, \quad SP_{rez} = \sum_i (y_i - \hat{y}_i)^2$$

cele trei sume de pătrate care apar în identitatea introdusă la definirea coeficientului de determinare. Sumele sunt referite ca suma pătratelor globală ($SP_g$), suma pătratelor datorate regresiei ($SP_{reg}$) şi suma pătratelor reziduale ($SP_{rez}$). Fiecare sumă de pătrate

are ataşat un număr de grade de libertate: $\nu_g = n\text{-}1$, $\nu_{reg} = p\text{-}1$, $\nu_{rez} = n\text{-}p$ şi se poate realiza un tabel al analizei dispersionale (ANOVA) sub forma

| Sursa de variație | Suma de pătrate | Grade de libertate | Media pătrată | F |
|---|---|---|---|---|
| **Regresie** | $SP_{reg}$ | $\nu_{reg}$ | $SP_{reg} / \nu_{reg} = s^2_{reg}$ | $F = s^2_{reg} / s^2$ |
| **Reziduală** | $SP_{rez}$ | $\nu_{rez}$ | $SP_{rez} / \nu_{rez} = s^2$ | |
| **Globală** | $SP_g$ | $\nu_g$ | $SP_g / \nu_g$ | |

**Testul F de semnificație globală**

Primul test utilizat în analiza regresiei este un test global de semnificație a ansamblului coeficienților (exceptând termenul liber, dacă acesta apare).

Ipotezele testului sunt

$H_0$: $\alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$

$H_1$: $(\exists)i$, astfel încât $\alpha_i \neq 0$.

În condițiile ipotezei nule, se demonstrează că statistica F, calculată în tabelul ANOVA, este repartizată Fisher-Snedecor $F_{p\text{-}1;n\text{-}p}$, încât se poate verifica ipoteza nulă.

Nerespingerea ipotezei nule duce la concluzia că datele observate nu permit identificarea unui model liniar valid, deci regresia nu este adecvată în scopul de prognoză, propus inițial.

**Teste t**

În situația când este respinsă ipoteza nulă, se acceptă că ecuația de regresie este semnificativă la nivel global, cu mențiunea că s-ar putea ca anumiți coeficienți să nu fie semnificativi. Pentru testarea fiecărui coeficient se utilizează un test t cu ipotezele:

$H_0$: $\alpha_i = 0$

$H_1$: $\alpha_i \neq 0$.

În condițiile ipotezei $H_0$ se arată că statistica $t_i = \dfrac{a_i}{s(a_i)}$ este repartizată Student cu $n - p$ grade de libertate, ceea ce permite utilizarea testului t. În expresia care dă statistica testului, $s(a_i)$ este abaterea standard estimată a coeficientului, dată ca rădăcina pătrată din elementul corespunzător de pe diagonala principală a matricei $s^2(\mathbf{X'X})^{-1}$.

Nerespingerea ipotezei nule arată că datele experimentale nu permit stabilirea necesității prezenței variabilei $x_i$ în model, variabila este nesemnificativă în model.

## Intervale de încredere

Apar de interes două tipuri de intervale de încredere: pentru parametrii modelului, $\alpha_i$, şi pentru valorile prognozate cu ajutorul modelului estimat.

**Parametrii modelului**

O regiune de încredere, la nivelul $\delta$, pentru ansamblul parametrilor este dată de

$$(\boldsymbol{\alpha} - \mathbf{a})'\mathbf{X'X}(\boldsymbol{\alpha} - \mathbf{a}) \leq ps^2 F_{1-\delta,p,n-p}$$

Utilizând repartiția statisticilor $t_i$, definite la testarea semnificației parametrilor, se demonstrează că **intervalul de încredere pentru parametrul $\alpha_i$**, $i = 1, 2, \ldots, p$, este dat la pragul de încredere $\alpha$, de relația

$$a_i - t_{1-\alpha/2;n-p} s(a_i) \leq \alpha_1 \leq a_i + t_{1-\alpha/2;n-p} s(a_i) .$$

**Valorile prognozate**

Utilitatea principală a modelului liniar este prognozarea valorilor variabilei dependente. Valoarea prognozată este evident o statistică pentru că se obține prin modelul estimat (din datele experimentale). Se poate atunci vorbi de repartiția de sondaj a valorii prognozate, repartiție care stă la baza determinării intervalelor de încredere pentru valorile prognozate.

În estimarea intervalului de încredere pentru o valoare $y_0 = \mathbf{x_0}\alpha + \varepsilon_0$, se distinge între situațiile în care observația $\mathbf{x_0}$ a fost, sau nu, utilizată la estimarea coeficienților (cu alte cuvinte, dacă matricea **X** conține sau nu linia $\mathbf{x_0}$).

În primul caz, intervalul de încredere pentru valoarea estimată este

$$\hat{y}_0 - t_{1-\alpha/2;n-p} s \sqrt{x_0 (X'X)^{-1} x_0'} \le y_0 \le \hat{y}_0 + t_{1-\alpha/2;n-p} s \sqrt{x_0 (X'X)^{-1} x_0'}$$

unde $\hat{y}_0 = \mathbf{x_0 a}$, este valoarea prognozată de ecuația de regresie.

În al doilea caz, intervalul de încredere este

$$\hat{y}_0 - t_{1-\alpha/2;n-p} s \sqrt{x_0 (X'X)^{-1} x_0' + 1} \le y_0 \le \hat{y}_0 + t_{1-\alpha/2;n-p} s \sqrt{x_0 (X'X)^{-1} x_0' + 1} \ .$$

În cazul regresiei simple (dreapta de regresie), ultimul interval de încredere are forma

$$\hat{y}_0 - t_{1-\alpha/2;n-p} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \le y_0 \le \hat{y}_0 + t_{1-\alpha/2;n-p} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \ ,$$

de unde se obține concluzia că valorile prognozate au intervale de încredere, la același prag de încredere, mai mari pe măsură ce valoarea $x_0$ este mai depărtată de media $\bar{x}$. De aici apare recomandarea ca un model liniar să nu fie utilizat pentru prognoză în cazul în care variabilele independente au valori depărtate de centrul datelor considerate la estimarea modelului (de exemplu, estimarea trendului ratei de schimb valutar din datele unei săptămâni nu poate fi utilizată pentru a prognoza rata de schimb de peste un an). În cazul unui sistem dinamic (valorile sunt produse/evaluate în timp), prognoza se va realiza doar pentru câteva momente de timp, după care are loc o nouă estimare a modelului etc.

## Analiza reziduurilor

Analiza statistică a ecuației de regresie este bazată pe ipotezele Gauss-Markov asupra erorilor $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$. Valabilitatea acestor ipoteze, în special cea a normalității erorilor, poate fi testată prin analiza reziduurilor. Ca şi în cazul testelor statistice, concluziile analizei sunt de genul: ipoteza normalității se respinge sau ipoteza normalității nu se respinge. Analiza reziduurilor este, în esență, de natură grafică.

Calculul estimațiilor erorilor produce

$$\mathbf{e = Y_{obs} - Y_{est} = Y_{obs} - Xa = Y_{obs} - X(X'X)^{-1}X'Y_{obs} = (1 - X(X'X)^{-1}X')Y_{obs}}$$

Notând $\mathbf{Z = X(X'X)^{-1}X'} = (z_{ij})$, rezultă că, în cazul îndeplinirii ipotezelor Gauss-Markov, dispersia reziduului $e_i$ este egală cu $(1-z_{ii})\sigma^2$ unde $z_{ii}$ sunt elementele de pe diagonala principală a matricei **Z**, cu estimația $s^2(e_i) = (1-z_{ii})s^2$. Reamintim că media reziduurilor este egală cu zero.

Ipotezele de repartiție a erorilor sunt reflectate în repartiția reziduurilor (estimații ale erorilor). Se analizează histograma reziduurilor sau diagrame ale reziduurilor în raport de valorile estimate, de variabilele independente. Diagramele construite în continuare pun în evidență eventualele abateri de la repartițiile presupuse pentru erori, abateri ce vor exprima deviațiile de la ipotezele de repartiție a erorilor.

**Diagrama reziduurilor**

Deoarece $e_i \sim N(0;(1-z_{ii})\sigma^2)$, rezultă că mărimile $d_i$, $i = 1,\ldots,n$, date de

$$d_i = \frac{e_i}{s\sqrt{1-z_{ii}}}$$

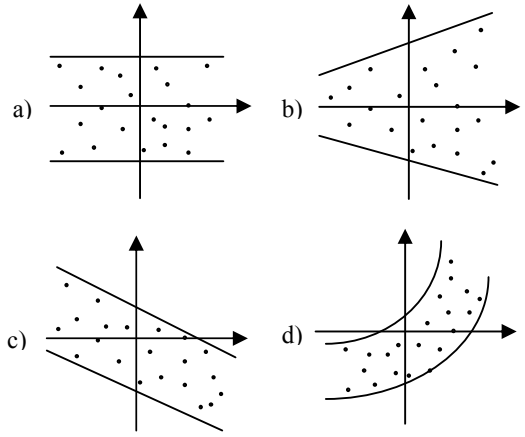
sunt repartizate N(0;1). Din acest motiv, mărimile $d_i$ sunt denumite reziduuri normalizate.

**Observație**. În practică, se neglijează uneori radicalul de la numitor.

Histograma mărimilor $d_i$ trebuie să reflecte o repartiție normală standard. Atunci când $n$ este relativ mic, histograma va prezenta, în general, mari neregularități față de situația care ar permite aproximarea cu o curbă normală. Decizia referitoare la proveniența, sau neproveniența, dintr-o repartiție normală se poate lua în acest caz, de exemplu, în urma comparației cu histograme obținute pentru eşantioane de acelaşi volum $n$ generate aleatoriu dintr-o repartiție normală standard.

**Diagrama reziduuri – valori estimate**

Considerând punctele de coordonate $(\hat{y}_i,d_i)$, $i = 1,\ldots,n$, reprezentate într-un sistem de axe rectangulare, sunt posibile 4 situații caracteristice, sau combinații ale lor, de regiuni ocupate de punctele considerate.




Cazul a) nu arată nici o abatere de la normalitate şi nici o violare a ipotezei că erorile au aceeaşi dispersie constantă.

În cazul b), se constată o creştere a dispersiei, deci este invalidată ipoteza constanței dispersiei erorilor. Practic, în această situație se consideră că modelul nu conține o variabilă esențială, cum ar fi timpul, sau că metoda de calcul adecvată este metoda celor mai mici pătrate ponderate. În anumite situații reale, situația poate fi rezolvată şi printr-o transformare prealabilă a datelor (de exemplu, prin logaritmare).

Cazul c) arată practic o eroare de calcul, deoarece este ca şi cum nu s-ar fi reuşit explicarea unei componente liniare a variației variabilei dependente.

Cazul al patrulea, d), arată că modelul nu este adecvat datelor observate. Se încearcă un nou model care să includă variabile de ordin superior, de genul $x^2$, care să preia variația curbilinie, sau se transformă în prealabil variabila y.

**Observație**. Indiferent de forma regiunilor, punctele foarte depărtate de celelalte oferă informații despre observațiile aberante. Regula uzuală este aceea ca orice observație pentru care $|d_i| > 3$ să fie considerată o observație aberantă. Practic, în acest caz, observațiile aberante se vor exclude din setul de date sau, dacă observațiile

sunt totuşi de interes, se va încerca obținerea unor determinări suplimentare în regiunea de interes. În ambele situații se va reface calculul regresiei.

**Diagrama reziduuri – variabilă independentă**

Se vor reprezenta grafic punctele de coordonate $(x_{ji}, d_i)$, $i = 1,\ldots,n$, pentru fiecare variabilă independentă $x_j$.

Cele patru situații grafice possibile se interpretează similar, cu observația că situația d) impune introducerea în model a variabilei $x_j$ ridicată la o putere.

## Multicoliniaritatea

Situația descrisă drept multicoliniaritate apare atunci când un grup de variabile independente sunt puternic corelate între ele. În acest caz, prin includerea în model a unei variabile din grup, restul variabilelor din grup nu mai aduc o informație semnificativă. Simultan are loc o supraevaluare a coeficientului de determinare, ca şi a dispersiilor coeficienților estimați, ceea ce poate denatura interpretarea modelului şi, în plus, produce mărirea intervalelor de încredere.

Apar astfel două probleme: determinarea multicoliniarității şi cum trebuie procedat în cazul existenței multicoliniarității.

**Detectarea multicoliniarității**

Cea mai simplă metodă de detectare a multicoliniarității este bazată pe studiul matricei de corelație dintre variabilele $x$. Se pot determina astfel perechile de variabile independente care sunt puternic corelate între ele. O structură mai complexă a intercorelațiilor poate fi detectată prin calcularea determinantului acestei matrice de corelație. O valoare apropiată de zero a determinantului reflectă o puternică corelație între anumite variabile, deci existența multicoliniarității.

O altă abordare a problemei este aceea a stabilirii unui indicator sintetic pentru a decide dacă o variabilă este coliniară cu celelalte (sau cu un grup dintre celelalte). Notând cu $R_i^2$ coeficientul de determinare obținut la estimarea regresiei multiple având ca variabilă dependentă pe $x_i$ şi ca variabile independente restul variabilelor $x$, adică

$$x_i = f(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$$

se introduce **toleranța** variabilei $x_i$ prin

$$\tau_i = 1 - R_i^2 .$$

O valoare mică a lui $\tau_i$ (uzual mai mică decât 0,1) reflectă un coeficient $R_i^2$ apropiat de 1, deci o legătură liniară puternică între $x_i$ şi restul variabilelor independente. Prin urmare $x_i$ este coliniară cu celelalte variabile independente.

Se defineşte **factorul de inflație a varianței**, notat *VIF*, inversul toleranței:

$$VIF_i = \frac{1}{\tau_i} .$$

Denumirea provine din aceea că un asemenea factor apare multiplicativ în definirea varianței coeficienților estimați (se poate spune că se măsoară de câte ori este supraevaluată varianța coeficienților datorită multicoliniarității în raport cu situația când nu ar exista coliniaritate). Interpretarea este dedusă din cea a toleranței: o valoare *VIF* mare (uzual mai mare decât 10), denotă coliniaritate.

**Eliminarea multicoliniarității**

O rezolvare comună a problemei multicoliniarității este aceea ca dintre două variabile independente corelate să se rețină în model doar una.

Prin interpretarea toleranțelor sau a factorilor de inflație se vor exclude din model acele variabile care au toleranțe mici (sau factori de inflație mari).

## Cea mai bună regresie

Procesul de selectare a celei mai bune regresii are loc în contextul în care există o variabilă dependentă *y* şi o mulțime de variabile independente posibile **x**. Problema poate fi formulată:

> *Care este acea submulțime minimală de variabile independente care permite estimarea unui model liniar semnificativ şi adecvat valorilor observate y?*

**Etapele selectării celei mai bune regresii**

1. Se identifică toate variabilele independente posibile (cu alte cuvinte se specifică modelul maxim).
2. Se specifică criteriul de selectare a celei mai bune regresii.
3. Se specifică o strategie pentru selectarea variabilelor independente.
4. Se realizează estimarea şi analiza modelului.
5. Se evaluează reliabilitatea modelului ales.

**Strategii de selectare a celui mai bun model**

### *Metoda tuturor regresiilor posibile*

Se estimează toate regresiile posibile.

Se reţin valorile coeficienților de determinare; gruparea este după cardinalul mulțimii de predictori.

| **Variabile independente** | **$R^2$** |
|---|---|
| $\{x_1\}$, $\{x_2\}$ … | … |
| $\{x_1,x_2\}$, $\{x_1,x_3\}$, …, $\{x_{n-1},x_n\}$ | … |
| … | … |
| $\{x_1,x_2,\ldots,x_n\}$ | … |

Se analizează valorile $R^2$ şi se reține acea submulțime de variabile pentru care se realizează compromisul acceptabil între numărul de variabile şi mărimea coeficientului de determinare.

### *Selecția prospectivă*

Procedura începe prin includerea în model a variabilei independente având cel mai mare coeficient de corelație cu variabila *y*. La fiecare pas următor, se analizează fiecare dintre variabilele neincluse încă în model printr-un test F secvențial şi se extinde modelul prin includerea acelei variabile care aduce o contribuție maximă (probabilitatea critică din testul F este cea mai mică). Procesul se opreşte atunci când modelul nu mai poate fi extins, criteriul uzual fiind acela al fixării un prag de intrare ($P_{IN}$) şi acceptând doar variabilele pentru care probabilitatea critică în testul F secvențial este mai mică sau egală cu acest prag.

Procedura are ca limitări faptul că anumite variabile nu vor fi incluse în model niciodată, deci importanța lor nu va fi determinată. Pe de altă parte, o variabilă inclusă

la un anumit pas rămâne permanent în model, chiar dacă, prin includerea ulterioară a altor variabile, importanța ei poate să scadă.

### *Selecția retrogradă*

Se începe cu estimarea modelului complet şi apoi, într-un număr de paşi succesivi, se elimină din model variabilele nesemnificative. La fiecare pas, pe baza unui test F parțial, se elimină acea variabilă care are cea mai mare probabilitate critică. Procesul se opreşte atunci când nici o variabilă nu mai poate fi eliminată. Criteriul uzual este acela de fixare a unui prag de eliminare ($P_{OUT}$) şi considerarea doar a variabilelor care au probabilitatea critică mai mare decât acest prag.

### *Selecția pas cu pas*

Procedura pas cu pas (*stepwise regression*) este o combinație a celor două metode descrise anterior. La un pas ulterior al regresiei prospective se permite eliminarea unei variabile, ca în regresia retrogradă. O variabilă eliminată din model devine candidată pentru includerea în model, iar o variabilă inclusă în model devine candidată la excludere. Pentru ca procesul să nu intre într-un ciclu infinit, trebuie ca $P_{IN} \leq P_{OUT}$.

## *B. Instrumente Excel, SPSS*

### Excel

### *REGRESSION*

Estimarea coeficienților unui model liniar prin metoda celor mai mici pătrate şi calculul statisticilor necesare testelor statistice asociate sunt efectuate de procedura **Regression**, una dintre cele mai complexe din pachetul de prelucrări statistice din Excel. Procedura permite şi construirea graficelor necesare pentru aprecierea vizuală a potrivirii modelului liniar. Deşi acestea, din motive evidente, necesită prelucrări suplimentare de scalare înainte de interpretare, existența lor este un real ajutor pentru statistician.

***Termeni***

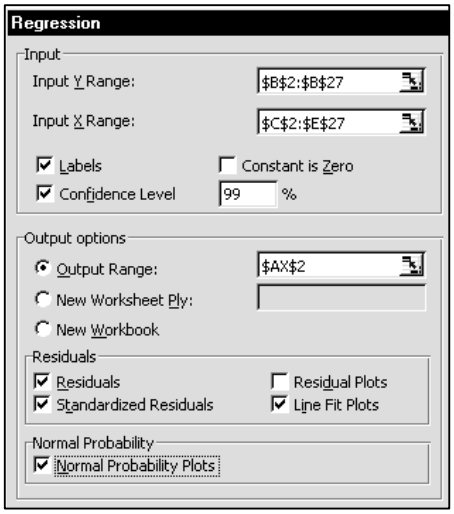
Modelul liniar estimat de procedură este

$$Y = \alpha_0 X_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_{p-1} X_{p-1} + \varepsilon,$$

care exprimă faptul că variabila Y se poate obține ca o combinație liniară a variabilelor $X_0, X_1, \ldots, X_{p-1}$ la care se adaugă o "eroare" $\varepsilon$.

Pentru estimarea parametrilor modelului se consideră disponibile *n* observații asupra tuturor variabilelor din model. Valorile sunt structurate ca un tablou dreptunghiular, fiecare variabilă ocupând o coloană (deci o linie este referită drept o observație).

Dialogul procedurii **Regression** este prezentat în figura următoare.



***Input***

*Input Y Range* – se precizează domeniul (coloana) pe care se află valorile variabilei dependente.

*Input X Range* – se precizează domeniul pe care se află valorile tuturor variabilelor independente. Acest domeniu trebuie să fie compact, fiecare variabilă $X_i$ ocupând o coloană.

*Labels* – se marchează boxa de control în cazul în care prima linie din tabloul de date este cu denumirile variabilelor (situaţie recomandată).

*Constant Is Zero* – se marchează boxa de control dacă modelul care se estimează este fără termen liber.

*Confidence Level* – se precizează, procentual, siguranţa statistică dorită în raportarea intervalelor de încredere deci valoarea $(1-\alpha)\times 100$, unde $\alpha$ este pragul de semnificaţie. Intervalele obţinute sunt suplimentare, întotdeauna afişându-se cele pentru $\alpha = 0{,}05$. Boxa se va marca doar dacă se doreşte şi un alt prag de semnificaţie.

***Output options***

*Output Range, New Worksheet Ply, New Workbook* – Precizează zona unde se vor înscrie rezultatele. Zona de rezultate este foarte complexă, cuprinde tabele care depind de mărimea modelului, de numărul de observaţii, de numărul graficelor dorite etc. Prin urmare se va prefera o foaie de calcul nouă sau o zonă liberă în dreapta şi în jos.

***Residuals***

*Residuals* – se marchează boxa de control în cazul când se doreşte calcularea reziduurilor modelului estimat.

*Residual Plots* – se marchează boxa de control în cazul când se doreşte obţinerea diagramelor *reziduuri – variabilă independentă,* adică vizualizarea punctelor de coordonate $(x_{ij}, r_j)$, $j = 1,\ldots n$, având ca abscisă o valoare a variabilei independente $X_i$, iar ca ordonată reziduul corespunzător.

*Standardized Residuals* – această boxă de control se va marca dacă se doreşte calculul valorilor standardizate ale reziduurilor. Valorile astfel obţinute provin, teoretic, dintr-o distribuţie normală standard, astfel încât o histogramă a acestor valori trebuie să se apropie de curba normală (clopotul lui Gauss).

*Line Fit Plots* – se marchează această boxă de control dacă se doreşte afişarea diagramelor *Y – variabilă independentă,* prin care se vizualizează, pe un acelaşi grafic, punctele de coordonate $(x_{ij}, y_{obs,i})$, $(x_{ij}, y_{est,i})$, $j = 1,\ldots,n$, unde abscisele sunt valorile variabilei independente, iar ordonatele sunt valorile observate şi cele estimate ale variabilei dependente. Este desenat câte un grafic pentru fiecare variabilă independentă.

Interpretarea acestor diagrame poate oferi indicaţii asupra adecvanţei modelului, asupra valorilor aberante.

***Normal Probability***

*Normal Probability Plots* – se marchează dacă se doreşte vizualizarea repartiţiei de sondaj a variabilei Y într-o reţea de probabilitate.

***Exemplu***

Un set de date cuprinde 25 de observaţii asupra a 4 variabile, notate Y (considerată variabila dependentă) şi $X_1$, $X_2$, $X_3$ (considerate variabile independente). Valorile şi denumirile ocupă în foaia de calcul un domeniu dreptunghiular continuu, B2:E27, valorile Y ocupând prima coloană.

Pentru a estima modelul liniar

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon,$$

cu termen constant, se apelează procedura **Regression**.

**a)** Un prim tabel de rezultate, prezentat în figura alăturată, conține statisticile generale ale ecuației de regresie.

| SUMMARY OUTPUT | |
|---|---|
| *Regression Statistics* | |
| Multiple R | 0.779 |
| R Square | 0.607 |
| Adjusted R Square | 0.550 |
| Standard Error | 2.023 |
| Observations | 25 |

*Multiple R* – coeficientul multiplu de corelație.

*R Square* – coeficientul de determinare (este egal cu pătratul coeficientului de corelație multiplă). Poate fi gândit, exprimat procentual, drept proporția din variația variabilei dependente explicată de variația variabilelor independente: 60,7% din variația lui Y este explicată de variabilele X.

*Adjusted R Square* – valoarea corectată a coeficientului de determinare. Este introdusă pentru a contracara (parțial) efectul creşterii mecanice a lui $R^2$ o dată cu numărul variabilelor independente.

*Standard Error* – eroarea standard a estimației. Se calculează ca abaterea standard a reziduurilor (pentru numărul gradelor de libertate utilizat se va vedea tabloul ANOVA, în continuare) şi este estimația abaterii standard a erorilor ε (în ipoteza normalității acestora).

*Observations* – numărul de observații din eşantion.

**b)** Al doilea tabel de rezultate cuprinde tabloul de analiză a varianței asociat regresiei estimate.

| ANOVA | | | | | |
|---|---|---|---|---|---|
| | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 3 | 132.476 | 44.159 | 10.791 | 0.0002 |
| Residual | 21 | 85.937 | 4.092 | | |
| Total | 24 | 218.414 | | | |

Coloanele acestui tablou au semnificațiile uzuale într-un tablou ANOVA:

Sursa de variație – arată descompunerea variației totale în variația explicată de regresie şi cea reziduală (neexplicată).

df – numărul gradelor de libertate: 3 = p – 1, 21 = n – p, 24 = n – 1, unde p = 4 este numărul parametrilor modelului (trei variabile X plus termenul liber) iar n = 25 este numărul de observații.

SS – sumele de pătrate potrivit descompunerii

$$\begin{matrix}\text{Suma globală}\\ \text{de pătrate}\end{matrix} = \begin{matrix}\text{Suma de pătrate}\\ \text{datorată regresiei}\end{matrix} + \begin{matrix}\text{Suma de pătrate}\\ \text{reziduală}\end{matrix}$$

MS – media sumelor de pătrate: SS împărțită la numărul respectiv de grade de libertate. Valoarea de pe linia a doua (*Residual*) este estimația dispersiei pentru repartiția erorilor şi este pătratul erorii standard a estimației.

F – valoarea statisticii F pentru testul caracterizat de

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ H_1 : \text{există cel puțin un coeficient } \alpha_i \text{ diferit de zero.} \end{cases}$$

Acest test se referă la ansamblul variabilelor independente (este de remarcat că $H_0$ nu se extinde şi asupra termenului liber). Datorită înțelesului ipotezei nule, se consideră că prin acest test se verifică semnificația întregii regresii.

*Significance F* – este probabilitatea critică unilaterală. Dacă valoarea afişată este mai mică decât pragul de semnificație fixat, atunci se respinge ipoteza nulă în favoarea ipotezei alternative.

**c)** Al treilea tablou de rezultate conține valorile estimate pentru coeficienții modelului, precum şi statisticile necesare verificării ipotezelor uzuale asupra coeficienților. De remarcat că, spre deosebire de testul F, testele asupra coeficienților sunt individuale.

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 11.718 | 5.421 | 2.161 | 0.042 | 0.444 | 22.992 |
| x1 | -1.443 | 0.398 | -3.623 | 0.002 | -2.271 | -0.615 |
| x2 | 3.135 | 0.593 | 5.289 | 0.000 | 1.902 | 4.367 |
| x3 | -0.324 | 0.169 | -1.919 | 0.069 | -0.675 | 0.027 |

**Liniile** tabelului se referă la variabilele din model, incluzând şi termenul liber.

**Coloanele** tabelului sunt următoarele:

(prima coloană) – sunt afişate denumirile existente în tabloul de date sau create automat pentru variabilele independente implicate. **Intercept** este denumirea pentru termenul liber (constant) al modelului.

*Coefficients* – conține valorile estimate ale coeficienților. Din valorile afişate rezultă că modelul estimat în exemplu este

$$Y = 11{,}718 - 1{,}443*X_1 + 3{,}135*X_2 - 0{,}324*X_3.$$

În ipotezele distribuționale ale modelului liniar, valorile calculate ale coeficienților provin din repartiții normale, fiind astfel posibile verificări statistice ale coeficienților.

*Standard Error* – eroarea standard a coeficientului (abaterea standard a repartiției coeficientului).

*t Stat* – statistica t pentru verificarea ipotezei $H_0 : \alpha_i = 0$ contra ipotezei alternative $H_1 : \alpha_i \neq 0$. În condițiile ipotezei nule se demonstrează că raportul dintre coeficient şi eroarea standard a coeficientului urmează o repartiție Student cu (n – p) grade de libertate. Acest raport este tocmai valoarea raportată drept t Stat. Adică 2,161 = 11,718/5,421 etc. Utilizarea statisticii este cea uzuală.

*P-value* – probabilitatea critică bilaterală a testului t cu ipotezele precizate la t Stat. Pentru pragul de semnificație $\alpha = 0{,}05$ se poate respinge ipoteza de nulitate a termenului liber (0,042 < 0,05) şi a coeficienților $\alpha_1$ şi $\alpha_2$ (0,002 şi 0,000 sunt mai mici decât 0,05). Nu se poate respinge ipoteza nulă privind coeficientul $\alpha_3$ (0,069 > 0,05).

*Lower 95%, Upper 95%* – limitele inferioară şi superioară ale intervalului de încredere pentru parametrul respectiv. Limitele la pragul 0,05 sunt calculate automat, indiferent de inițializarea procedurii Regression.

Se poate deci interpreta că, în populație, parametrii modelului liniar sunt cuprinşi în intervalele următoare:

$$0{,}444 < \alpha_0 < 22{,}992$$
$$-2{,}271 < \alpha_1 < -0{,}615$$
...

Se poate observa că ultimul interval cuprinde şi valoarea zero, prin urmare se regăseşte concluzia privind nerespingerea ipotezei nule $H_0 : \alpha_3 = 0$.

**d)** Studiul reziduurilor se poate face pe baza datelor raportate în tabelul alocat reziduurilor, tabel având structura următoare:

RESIDUAL OUTPUT

| Observation | Predicted y | Residuals | Standard Residuals |
|---|---|---|---|
| 1 | 11.38 | -2.05 | -1.08 |
| 2 | 10.16 | -4.44 | -2.34 |
| ... | ... | ... | ... |
| 25 | 12.85 | 1.25 | 0.66 |

Pentru fiecare observație (linie din tabelul de date inițial) se afişează:

*Observation* – numărul de ordine al observației.

*Predicted y* – valoarea y prognozată pentru observația respectivă; se obține înlocuind valorile X ale observației în modelul estimat.

*Residuals* – valoarea erorii de predicție (diferența dintre valoarea observată şi valoarea prognozată).

*Standard Reziduals* – valoarea standardizată a erorii. Este obținută prin împărțirea reziduului la abaterea standard a reziduurilor (rezultatul nu este susținut absolut riguros de teorie).

**e)** Analiza calității modelului este facilitată şi de graficele construite automat de procedura **Regression**. Sunt produse două tipuri de diagrame:

- diagrame reziduuri vs. variabile independente şi
- diagrame variabila dependentă vs. variabile independente.

Graficele necesită, de obicei, prelucrări suplimentare pentru a fi interpretate sau raportate.

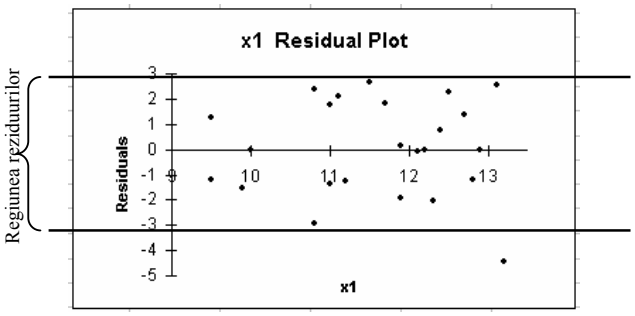



Diagrama reziduuri – variabilă

În figură se dă un exemplu de diagramă reziduuri – variabilă independentă X. Punctele din figură se pot considera într-o regiune de tip bandă orizontală ceea ce nu contrazice ipotezele de normalitate a erorilor. Forma de bandă uniformă reflectă constanța dispersiei reziduurilor pentru tot domeniul variabilei independente $X_1$. Alte forme de distribuire a reziduurilor duc la concluzii importante pentru adecvanța modelului în privința variabilei independente implicate:

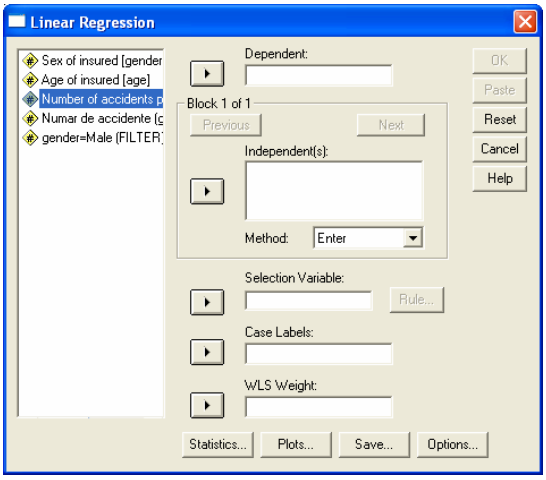
| Forma regiunii | Interpretare |
|---|---|
| | Situația "bună". Nu se contrazic ipotezele de normalitate făcute asupra erorilor. |
| | Dispersia erorilor nu este constantă (se modifică după valorile X). Se poate ca din model să fie omisă o variabilă de gen "Timp". |
| | Modelul liniar nu este adecvat în privința variabilei independente respective. Se poate încerca un introducerea unui termen pătratic. |
| | Situația poate să apară în urma unei erori de calcul. Practic ar însemna că nu s-a considerat componenta liniară, adică scopul modelului nu a fost atins. |

În mod asemănător se pot interpreta diagramele Y – X.

## SPSS

### *Dreapta de regresie*

Principalul dialog pentru estimarea unui model liniar se obține prin **Analyze – Regression –Linear.**

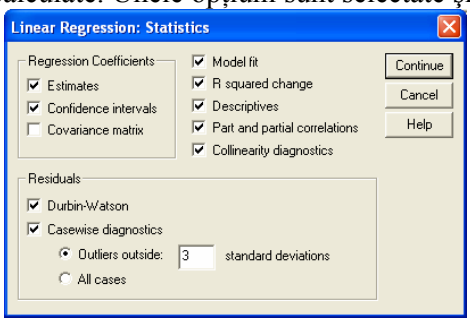



În ***Dependent*** se va transfera variabila dependentă. Variabilele independente, ***Independent****(s)*, pot fi grupate pe blocuri: 1. se transferă variabilele dorite, 2. se precizează în *Method* modul de introducere a acestor variabile în regresie (*Enter* – toate simultan, *Forward*, *Backward*, *Stepwise* – metodele discutate la alegerea celei mai bune regresii), 3. se defineşte un nou bloc prin *Next*.

Se pot selecta observaţiile precizând în ***Selection Variable*** variabila şi, prin **Rule**, regula de selectare a cazurilor în funcţie de valorile variabilei de selecţie.

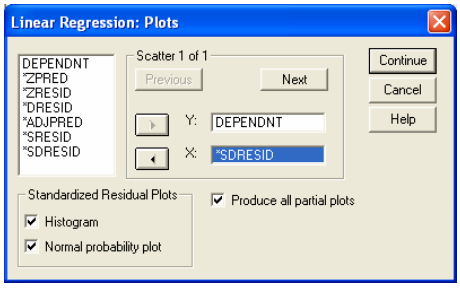
În ***Case Labels*** se poate preciza variabila care identifică cazurile, etichetele fiind considerate la reprezentările grafice. Prin ***WLS Weight*** se poate preciza variabila de ponderare pentru metoda celor mai mici pătrate ponderate (nediscutată în curs).

Butonul **Statistics** deschide dialogul sinonim în care se pot preciza statisticile calculate. Unele opţiuni sunt selectate şi în mod implicit.




*Estimates* – coeficienţii estimaţi, *Confidence intervals* – intervalele de încredere ale coeficienţilor, *Model fit* – calcularea statisticilor R, $R^2$ şi a tabelului ANOVA, *R squared change* – modificărea coeficientului de determinare şi testarea semnificaţiei schimbării la adăugarea fiecărui bloc de variabile, *Descriptive* – statisticile esenţiale pentru fiecare variabilă, *Collinearity diagnostics* – calcularea toleranţelor, a statisticilor VIF şi studiul multicoliniarităţii prin analiza în componente principale (a se vedea capitolul următor al cursului). În zona ***Residuals*** se produce o analiza a reziduurilor pentru a putea decide asupra normalităţii acestora şi a diagnostica valorile aberante.

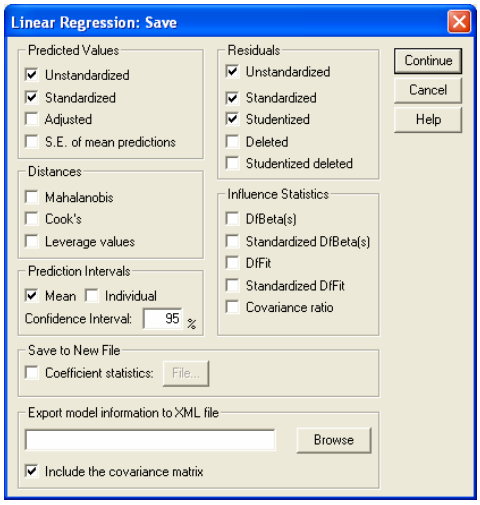
Prin **Plots** se afişează dialogul sinonim în care se pot indica reprezentările grafice dorite.




În lista variabilelor disponibile pentru diagrame se află DEPENDNT – variabila dependentă – şi variabile derivate din regresie cum ar fi valorile prognozate standardizate (*ZPRED), reziduurile standardizate (*ZRESID).

Diagramele indicate în *Standardized Residual Plots* sunt utile pentru verificarea normalităţii reziduurilor.

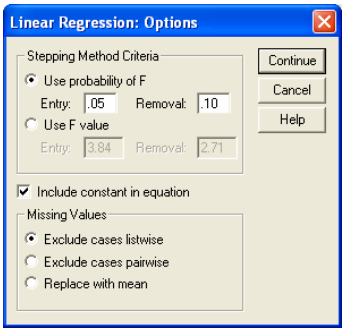
Dialogul **Save** permite calcularea şi salvarea ca variabile noi a valorilor prognozate şi a reziduurilor sub diferite forme, precum şi salvarea altor statistici de interes. ***Predicted Values*** – valorile prognozate prin model pentru fiecare caz: *Unstandardized*, *Standardized* pentru valorile nestandardizate şi standardizate, *Adjusted* valoarea prognozată pentru un caz din ecuaţia de regresie estimată fără a considera acel caz, *S.E. of mean predictions* abaterile standard ale valorilor prognozate, utile pentru calcularea intervalelor de încredere ale acestor valori.




***Distances*** – distanţele cazurilor de la punctul mediu, pentru identificarea valorilor aberante: *Mahalanobis* este distanţa explicată în capitolul privind clasificarea, *Cook's* este măsura a cât de mult se modifică reziduurile dacă se elimină cazul respectiv din estimarea modelului (o valoare mare arată o influenţă considerabilă a cazului în estimarea coeficienţilor), *Leverage values* măsoară influenţa cazurilor în estimare.

***Prediction Intervals*** sunt intervalele de incredere pentru valorile estimate, la nivelul de încredere precizat în *Confidence Interval.* Sunt generate două variabile.

***Residuals*** – reziduurile estimării în diferite forme: standardizate, nestandardizate, studentizate (reziduul este împărţit la estimaţia abaterii sale standard, proprie fiecărui caz). *Deleted*, *Studentized deleted* se referă la reziduurile obţinute din modelul la estimarea căruia cazul respectiv a fost exclus.




***Influence Statistics*** sunt modificările în coeficienţi (inclusiv cei standardizaţi), *DfBeta(s)* şi *Standardized DfBeta*, şi în valorile prognozate, *DfFit* şi *Standardized DfFit*, rezultate după excluderea cazului din estimare.

În sfârşit, prin butonul **Options** se deschide dialogul sinonim în care se pot fixa parametri ai estimării: pragurile de intrare şi excludere la metodele pas cu pas precum şi modul de tratare a valorilor lipsă dintr-o variabilă implicată.

## C. Lucrarea practică

**1.** Legea lui Ohm, I = V/R, afirmă că intensitatea curentului, I, este proporţională cu tensiunea, V, şi invers proporţională cu rezistenţa, R. Elevii dintr-un laborator de fizică efectuează experimente bazate pe legea lui Ohm: variază tensiunea, măsoară intensitatea curentului şi determină în final rezistenţa firului. Se obţin rezultatele:

| | | | | | |
|---|---|---|---|---|---|
| **V** | 0,50 | 1,00 | 1,50 | 1,80 | 2,00 |
| **I** | 0,52 | 1,19 | 1,62 | 2,00 | 2,40 |

Deoarece legea lui Ohm poate fi rescrisă sub forma unei regresii liniare, $I = \alpha+\beta V$, unde $\alpha = 0$ şi $\beta=1/R$, să se estimeze, pe baza datelor experimentale, coeficienţii $\alpha$ şi $\beta$.

– Să se obțină intervalul de încredere, la pragul de semnificație de 5%, pentru coeficientul $\beta$. Să se deducă intervalul de încredere pentru rezistența firului.
– Să se verifice ipoteza $\alpha = 0$.

**2.** O familie înregistrează consumul de gaz necesar încălzirii locuinței. Consumul (în mc) este raportat în tabelul următor, împreună cu diferența medie de temperatură față de cea externă (în grade Fahrenheit).

| Luna | oct | nov | dec | ian | feb | mar | apr | mai | iun |
|---|---|---|---|---|---|---|---|---|---|
| temperatura | 15.6 | 26.8 | 37.8 | 36.4 | 35.5 | 18.6 | 15.3 | 7.9 | 0 |
| Gaz | 520 | 610 | 870 | 850 | 880 | 490 | 450 | 250 | 110 |

– Să se studieze forma relatiei dintre cei doi indicatori. Exista asociere între cei doi indicatori?
– Să se estimeze dreapta de regresie care modelează relația dintre cei doi parametri.
– În timpul verii, proprietarul locuinței îmbunătățeşte izolația termică a casei sale. Drept care în luna februarie următoare, la o diferență medie de 40, se consuma 895 mc de gaz. Se poate spune că lucrarea efectuată reduce consumul de gaz?

**3.** Datele necesare acestul exercițiu sunt la adresa web www.infoiasi.ro/~val/statistica/boston.sav şi sunt doar o oglindire a unor date din surse internaționale. Analiza datelor doreşte să prognozeze prețul de vânzare a unei case din regiunea Boston în funcție de caracteristici diverse ale locuinței şi ale localizării ei. Prelucrarea se va efectua, de preferință, în SPSS

Variabilele sunt în ordine: CRIM – rata criminalității, ZN – proporția teritoriului zonat în loturi de peste 25,000 sq.ft., INDUS proporția terirorială a zonei industriale, CHAS – indicator de învecinare cu râul din zonă (= 1 da, 0 nu), NOX – concentrația de oxizi nitrici, RM – numărul mediu de camere, AGE – proporția de locuințe construite înainte de 1940 şi ocupate de proprietar, DIS – distanța ponderată la cinci centre din Boston, RAD – indicele de accesibilitate la rețeaua de autostrăzi, TAX – rata de impozit (procent la 10000$), PTRATIO – raportul copii-profesori în zonă, B – $1000(Bk-0.63)^2$ unde Bk este procentajul populației de culoare în zonă, LSTAT – procentajul populației sărace, MEDV – valoarea medie a caselor (în mii de dolari).

Se se efectueze următoarele operații:
– Completați în SPSS denumirile de variabile şi informațiile necesare.
– Verificați condițiile necesare aplicării analizei regresionale.
– Estimați ecuația de regresie prin diferite metode. Analizați dacă obțineți un răspuns care pare consistent, independent de metodă.
– Validați şi interpretați rezultatele regresiei.