Folosirea pachetului MathCad în analiza statistică

George Daniel Mateescu[*]
Corina Saman[*]
Mihai Buneci[*]

Rezumat

MathCad-ul ne oferă posibilitatea de a face o analiză a datelor şi de a testa modele posibile într-un mod simplu şi uşor de folosit dacă cunoastem matematica care se afla în spatele modelului.
Putem verifica ipotezele modelului regresiei cum ar fi dispersia constantă a erorilor, testarea ipotezelor statistice față de coeficienți, corelația erorilor, normalitatea erorilor ca variabile aleatoare.
Cuvinte cheie: regresie, ipoteze statistice
JEL: **C22**(Time-Series Models), **C51**(Model Construction and Estimation)

Pachetul **Mathcad** intră în categoria mediilor soft care asistă utilizatorul specialist într-un anumit domeniu. Permite realizarea de calcule matematice foarte complexe precum şi activități conexe cum ar fi reprezentări grafice, realizarea unei documentații, precum şi posibilitatea de export spre alte medii Windows.

Comenzile principale:

- **File**: operații uzuale privind fişierul ca entitate, deschidere, salvare, tiparire, transmitere prin fax sau poştă electronică
- **Edit**: comenzi de "dactilografiere inteligentă": ştergere, copiere, etc.
- **View**: se referă în special la subseturile de comenzi dedicate pentru diferite operații matematice şi care sunt grupate în seturi (palete) ce pot fi activate prin ***paleta matematică***
- **Insert**: se referă la obiecte matematice complexe cum sunt grafice de funcții, tabele (matrici) funcții (sistemul conține în formă predefinită cele mai cunoscute funcții matematice)
- **Format**: se referă atât la indicații de "înfățişare" a lucrării cât şi la formate specifice calculelor matematice (cum este de exmplu numărul de zecimale care se afişează)
- **Math**: priveşte comanda de începere a calculelor dar şi indicații referitoare la unitățile de măsură
- **Symbolics**: reprezintă un element de noutate în raport cu calculele numerice şi conține elemente de inteligență artificială care permit evaluări simbolice (primitive sau derivare simbolică)

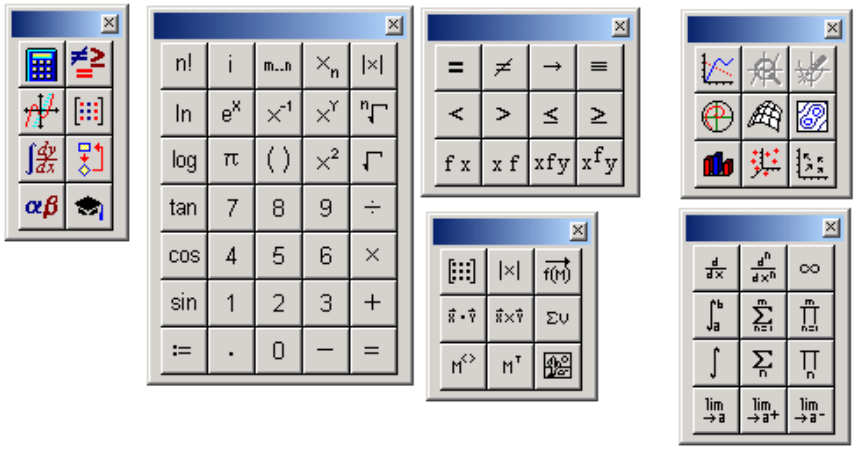
[*] Institutul de Prognoză Economică, Academia Romană

- **Window**: permite gestiunea mai multor lucrări
- **Help**: mediu complex care conține indicații de utilizare

Detalierea comenzilor. Tot ceea ce ține de tratarea fișierului ca entitate, vizualizare, tipărite editare, ferestre, help, sunt foarte asemănătoare cu alte pachete din mediul Windows (Word, Excel, etc.)

Paleta matematică, conține cele mai semnificative comenzi pe următoarele grupe

⇒ artimetică
⇒ logică
⇒ reprezentări grafice
⇒ matrici şi vectori
⇒ calcul diferențial şi integral
⇒ elemente de programare
⇒ alfabetul grec
⇒ evaluarea simbolică



Principiul de utilizare este al unei "foi de hârtie inteligente" în care utilizatorul scrie expresiile matematice în mod obişnuit iar "foaia" efectuează calculele şi pune rezultatele exact în continurea expresiei. Utilizatorul completează liniile şi coloanele matricii a iar apoi scrie pur şi simplu $a^{-1}=$ iar "foaia de calcul" pune rezultatul. Se pot face calcule oricât de complicate aşa cum se vede alăturat. Modul de introducere al expresiilor este "free hand"

$$a := \begin{bmatrix} 1 & 3 & 1 \\ 2 & 1 & -1 \\ 4 & 1 & 2 \end{bmatrix} \qquad b := \begin{bmatrix} 4 & 2 & 1 \\ 1 & 2 & 2 \\ 8 & 5 & 0 \end{bmatrix}$$

$$|a| \cdot b - a^{T} \cdot b^{-1} = \begin{bmatrix} -93.158 & -47.421 & -22.368 \\ -24.316 & -45.842 & -45.737 \\ -186.526 & -114.737 & 1.105 \end{bmatrix}$$

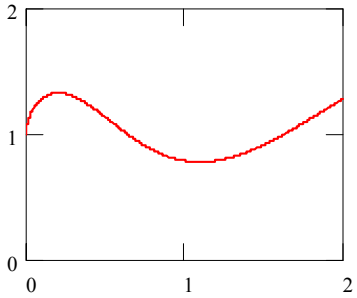
Calcule simbolice

$$\frac{d}{dx}\frac{\sin(x)}{x^3 - \cos\left(1+\sqrt{x}\right)} \rightarrow \frac{\cos(x)}{\left(x^3 - \cos\left(1+\sqrt{x}\right)\right)} - \frac{\sin(x)}{\left(x^3 - \cos\left(1+\sqrt{x}\right)\right)^2} \cdot \left(3 \cdot x^2 + \frac{1}{2} \cdot \frac{\sin\left(1+\sqrt{x}\right)}{\sqrt{x}}\right)$$

Reprezentări grafice

Se plasează:
pe orizontală, jos: variabila şi intervalul (domeniul e definiţie)
pe verticală, în stânga, legea de coespondenţă şi codomeniul
Mediul Windows este atât de puternic integrat încât permite, de exemplu în Word, editarea direct în pagină prin lansarea implicită în background a pachetului Mathcad!

$$\sqrt{x} + \frac{\cos(2 \cdot x)}{1 + x^2}$$



Calcule numerice

Integrala definită a unei funcţii a cărei primitivă nu este exprimabilă prin funcţii elementare. În cazul funcţiilor ale căror primitive se exprimă prin funcţii elementare se poate apela şi la calculul simbolic

$$\int_1^2 e^{-x^2}\,dx = 0.135$$

Sisteme de ecuaţii

Sistemele de ecuaţii liniare se rezolvă direct prin calcule cu matrici dar se pot rezolva şi ecuaţii sau sisteme de ecuaţii neliniare. Este necesară introducerea unor valori iniţiale pentru *x* şi *y* (semnificaţia acestei introduceri poate fi explicată numai după ce vor fi însuşite cunoştinţe de analiză numerică). Este posibil ca, uneori, să nu fie obţinută soluţia sistemului daca valorile de start nu satisfac anumite condiţii.
În exemlul alăturat semnul **=** se obţine prin selectare din tabelul de operatori de tip logic. Soluţia se obţine prin apelul funcţiei find care are exact acest rol.

$x := 4 \qquad y := 2$

given

$x^2 + 3\,y + x = 5$

$2 + x \cdot y = 3$

$\text{find}(x, y) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Analiză statistică

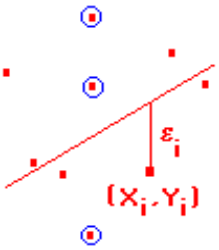
MathCad-ul ne oferă posibilitatea de a face o analiză a datelor şi de a testa modele posibile într-un mod simplu şi usor de folosit dacă cunoaştem matematica care se afla în spatele modelului.
Estimarea, deducţia şi predicţia sunt posibile pentru orice model simulat.
Cînd experimentăm modele posibile asociate fenomenelor economice trebuie luate decizii în funcţie de rezultatele obţinute în testarea ipotezelor statistice. In MathCad este suficient să formulăm ipoteza statistică, apoi măsurarea coincidenţei dintre valorile observate şi cele pe care ipoteza statistică le validează ca adevărate se face simplu calculînd statistica asociată ipotezei. Este necesar numai să traducem în formula potrivită ipoteza statistică. Avem un test statistic care este o funcţie a datelor observate. Tinînd cont că ipoteza nulă determină distribuţia de probabilitate a acelui test statistic care apoi determină probabilitate este simplu de construit orice test satistic.
In cele ce urmează vom explora posibilităţile de analiză a validităţii unui model de regresie liniară pe care ni le ofera MathCad-ul. Si acestea sunt: estimatori după metoda celor mai mici pătrate, verificare ipotezelor modelului liniar, deducţii asupra parametrilor funcţiei liniare, măsura potrivirii modelului pentru datele observate, analiza tabelei varianţei pentru regresie.
Datele folosite în exemplele de mai jos sunt: producţia industrială pi (variatie lunară,%), somajul somaj (rată lunară,%) si inflaţia inf (rată lunară,%) pentru perioada ianuarie 2000 - ianuarie 2002.

Conditiile impuse de Modelul regresiei Liniare

1. Variabila dependenta Y este o functie liniara de X.
2. Variabila eroare ε este componenta aleatoarea a modelului linear.
   Valorile lui X sunt presupuse fixe.
3. Termenii eroare corespunzatori obsevatiilor sunt necorelati. In plus, pentru orice valori date pentru X, erorile sunt variabile aleatoare normal distribuite cu media zero si dispersie constant

Cautam cei mai buni estimatori pentruintersectia cu axa Ox ($\beta_{0)}$ si panta dreptei ($\beta_{1)}$, in cazul modeluluiregresiei liniare simple dat prin ecuatia:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

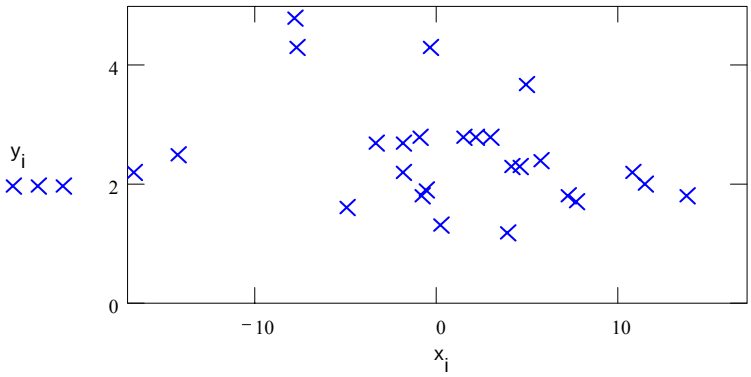


Termenul ε_i pentru observatia, i, este diferenta masurata pe verticala, dintre punctele observate (X_i, Y_i) si linia de regresie . Valuarea lui ε va fi positiva cind valorile observate se gasesc deasupra liniei de regresie si negativa cind se gasesc sub linie.

Un exemplu: regresia inflatiei(inf) fata de productia industriala(pi)
Reprezentarea grafica:

y := inf x := pi

n := rows (pi) i := 0 .. n − 1



Cum gasim aceasta dreapta?
Prin metoda celor mai mici patrate.

Suma patratelor Residualelor, SSE

Vom folosi **suma patratelor rezidualelor**

$$SSE(b_0, b_1) := \sum_i \left[y_i - (b_0 + b_1 \cdot x_i) \right]^2$$

Putem obtine o solutia folosin metoda de rezolvare a sistemelor de ecuatii in Mathcad

Valori initiale: $b_0 := 1 \quad b_1 := 1$

Blocul ecuatiilor

Given

$n \cdot b_0 + b_1 \cdot (\sum x) = \sum y$

$b_0 \cdot (\sum x) + b_1 \cdot \left(\sum \vec{x^2}\right) = \sum \left(\overrightarrow{x \cdot y}\right)$

$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} := \text{Find}(b_0, b_1)$

produce estimatorii pentru β_0 si β_1

$b_0 = 2.525 \qquad b_1 = -0.04$

pentru cele mai mici posibile sume de patrate,

$\text{SSE}(b_0, b_1) = 18.58$

$\text{SSE}(b_0 - 0.01, b_1 - 0.01) = 18.717$

Metoda celor mai mici patrate produce estimatorii cei mai buni - the **best linear unbiased estimators** (BLUE) - pentru coeficienti. Asta inseamna ca au dispersia minima.

In Mathcad sunt dati de functiile: intercept (x , y) slope (x , y)

intercept (x , y) = 2.525 slope (x , y) = −0.04

unde primul argument reprezinta variabila independenta. Spre deosebire de functia de corelatie ordinea argumentelor este esentiala !

intercept (x , y) ≠ intercept (y , x)

slope (x , y) ≠ slope (y , x)

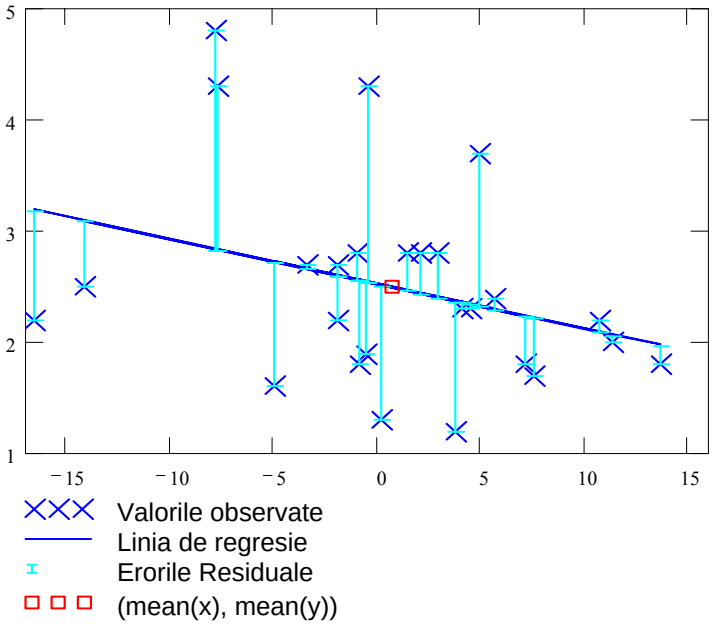
Linia de regresie pentru exemplu dat are ecuatia

$y_{lin}(x) := b_0 + b_1 \cdot x$

se vede pe grafic si de asemenea punctele observate si rezidualele. Punctul dat de valorile medii ale variabilelor aleatore:

(mean (x) , mean (y))

(indicat de un patrat) este de asemenea pe linia de regresie.



Estimatori pentru Dispersia Erorilor

Suma patratelor rezidualelor *SSE* poate fi folosita in estimarea dispersiei erorilor σ^2. Aceasta dispersie masoara imprastierea datelor observate fata de linia de regresie.
Un estimator pentru σ^2 in cazul unui esantion de lungime n este **mean square error** (*MSE*), care se calculeaza impartind valuarea lui *SSE* prin (n - 2)

$$MSE := \frac{SSE(b_0, b_1)}{n - 2} \qquad MSE = 0.774$$

Aceasta definitie urmareste ideea ca dispersia este media sumei patratelor deviatiei fata de linie. Impartim prin (n - 2) deoarece am estimat deja cei doi parametrii β_0 si β_1. Un estimator pentru abaterea standard a erorilor σ este **standard error of estimate**

$$se\_e := \sqrt{MSE} \qquad se\_e = 0.88$$

$$x := pi \qquad y := inf$$

Erorile reziduale, e, sunt date de:

$$y_{lin}(x, y) := intercept(x, y) + slope(x, y) \cdot x$$

$$e := y - y_{lin}(x, y)$$

Si estimatorul pentru abaterea standard este deci

$$se\_e(e) := \sqrt{\frac{\left(\sum \overrightarrow{e^2}\right)}{rows(e) - 2}} \qquad se\_e(e) = 0.88$$

Reziduale Standardizate

$$standard\_e(e) := \frac{e}{se\_e(e)} \qquad standard\_e(e)_1 = 0.119$$

Reziduale Studentizate

Un alt estimator pentru abaterea standard s a unei variabile X este:

$$i := 0 .. n - 1 \qquad leverage_i := \frac{1}{rows(x) - 1} \cdot \frac{\left(x_i - mean(x)\right)^2}{Var(x)} \qquad leverage_0 = 0.055$$

Ajustind rezidualele standardizate obtinem **reziduale studentizate**

$$student\_e_i := \frac{\frac{e_i}{se\_e(e)}}{\sqrt{\left[1 - \left(\frac{1}{rows(x)} + leverage_i\right)\right]}}$$

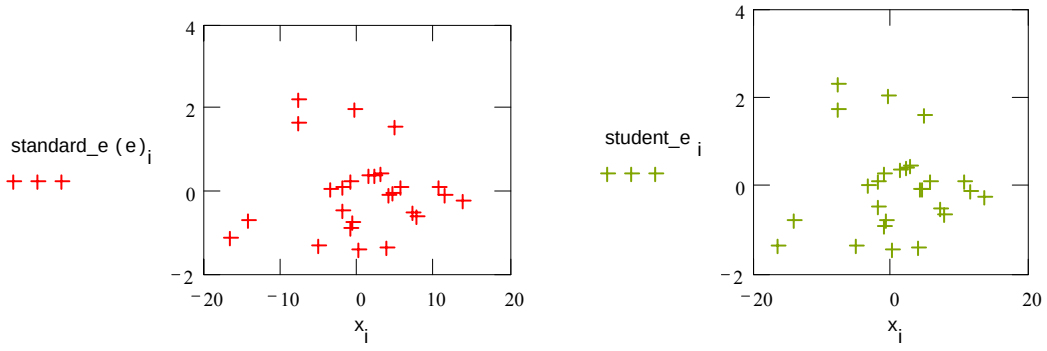
unde membrul drepteste un estimator pentru abaterea standard a rezidualelor e_i.

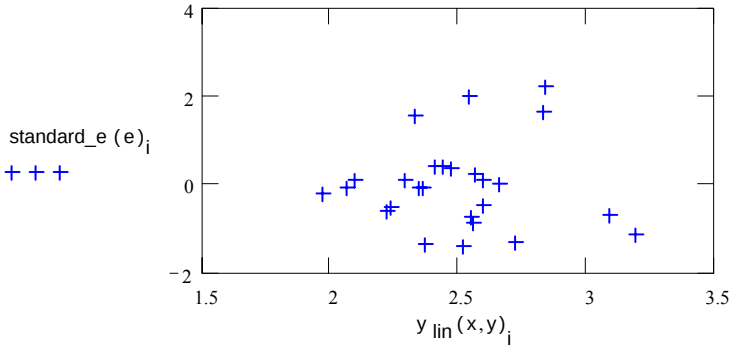
Rezidualele studentizate sunt mai precise in sensul ca indica mai exact diferentele in dispersia erorilor .

$$standard\_e(e)_1 = 0.119 \qquad student\_e_1 = 0.127$$

Graficul Rezidualelor

Daca datele urmeaza o relatie liniara atunci graficul rezidualelor standardizate (sau studentizate axa (y-axis) fata de valorile X sau fata de valorile previzionate de model nu vor prezenta nici un model evident ci vor fi repartizate aleator.





Daca datele nu respecta modelul liniar reziduale vor prezenta grafic un model(pattern) . De exemplu daca generam aleator un esantion de 100 valori dintr-o o serie de date care urmeaza o functie exponentiala:

$n\_esantion := 100$ $nr := 20$

$k := 0 .. n\_esantion - 1$ $x_{exp_k} := rnd(nr)$

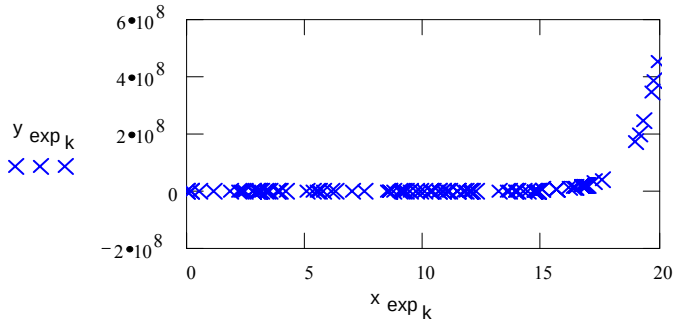
Si de asemenea erori normale cu dispersia σ:

$\sigma := 10$ $\varepsilon := rnorm(n\_esantion, 0, \sigma)$

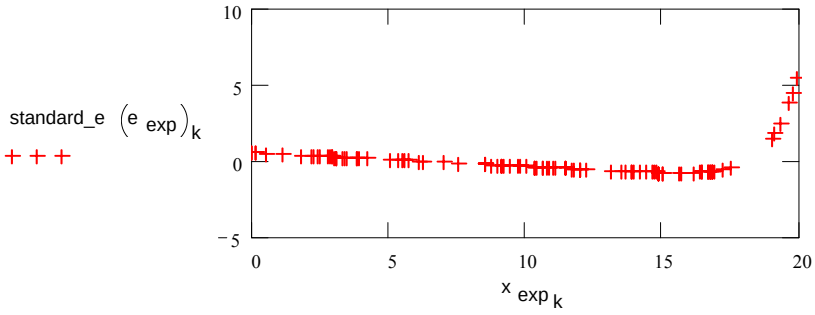
$y_{exp} := \overrightarrow{(exp(x_{exp}))} + \varepsilon$

$e_{exp} := y_{exp} - y_{lin}(x_{exp}, y_{exp})$

obtinem urmatorul grafic



si rezidualele



Dispersia Erorilor Constanta

Daca dispersia erorilor nu este constanta de la o valoare a lui X la alta , graficul rezidualelor va arata o repartizare crescatoare sau descrescatoare de la stinga la dreapta.
Sa vedem un exemplu de care urmeaza o relatie liniara asa cum arata graficul lui *Y* fata de *X* si cum arata si coeficientul de corelatie:

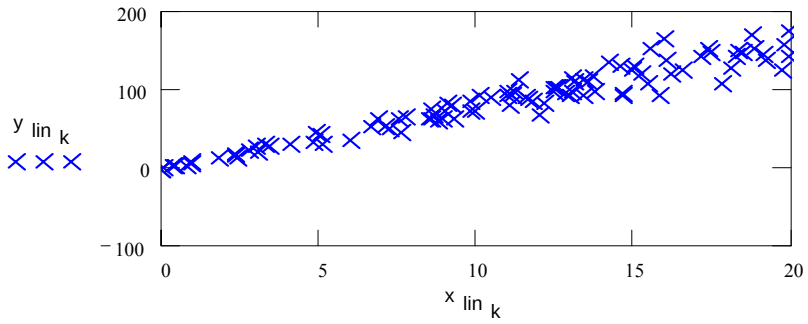
$n\_esantion := 100$ $nr := 20$

$k := 0 .. n\_esantion - 1$ $x_{lin_k} := rnd(nr)$

$\sigma := x_{lin}$ $\varepsilon_k := rnorm\left(n\_esantion, 0, \sigma_k\right)_k$

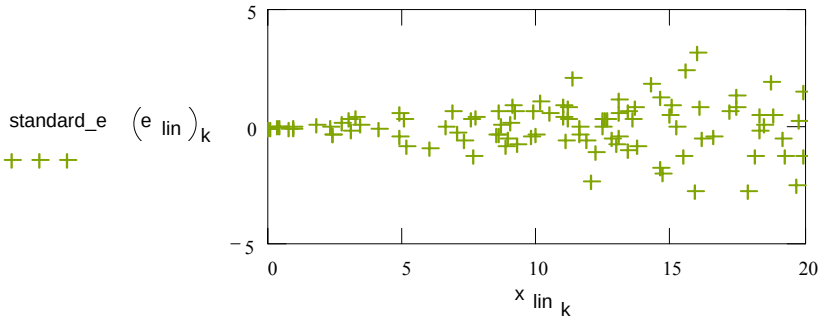
$m := 8$ $b := -2$ $y_{lin} := m \cdot x_{lin} + b + \varepsilon$

$e_{lin} := y_{lin} - y_{lin}\left(x_{lin}, y_{lin}\right)$



$corr\left(x_{lin}, y_{lin}\right) = 0.966$

dispersia erorilor prezinta o tendinta de crestere de la stinga la dreapta .



Corelatia Erorilor

Putem verifica daca exista corelatie intre termenii alaturati in seria erorilor folosind statistica **Durbin-Watson** (DW).

$e := y - y_{lin}(x, y)$

$$DW(e) := \frac{\sum_{d=1}^{rows(e)-1} \left(e_d - e_{d-1}\right)^2}{\sum_{d=0}^{rows(e)-1} \left(e_d\right)^2}$$

$DW(e) = 1.867$

Valori pentru statistica Durbin-Watson mai mici decit 2 indica corelatie positiva pentru erori si valori mai mari decit 2 o corelatie negativa. Totusi aceasta statistica nu poate da un raspuns decit pentru corelatia termenilor alaturati.

R^2 masura pentru dependenta cauzala intre variabila independenta si cea dependenta

Dependenta dintre variabile se poate exprima prin covarianta

$$E\left[\left(X-\mu_x\right)\cdot\left(Y-\mu_y\right)\right]$$

care de obicei se calculeaza normalizat prin coeficientul de corelatie ρ,

$$\rho=\frac{E\left[\left(X-\mu_x\right)\cdot\left(Y-\mu_y\right)\right]}{\sigma_x\cdot\sigma_y}$$

$$\rho_{pi.inf}:=\frac{\left[\sum_{i=0}^{rows(pi)-1}\left[\left(pi_i-\mu_{pi}\right)\cdot\left(inf_i-\mu_{inf}\right)\right]\right]\cdot\frac{1}{n}}{\sigma_{pi}\cdot\sigma_{inf}}$$

$\rho_{pi.inf}=-0.318$ $\quad corr(pi,inf)=-0.318$ $\quad rows(pi)=26$

$y:=inf$ $\quad x:=pi$

Erorile reziduale,*e*, sunt date de:

$y_{lin}(x,y):=intercept(x,y)+slope(x,y)\cdot x$

$e:=y-y_{lin}(x,y)$

Si estimatorul pentru abaterea standard este deci $se\_e(e):=\sqrt{\frac{\left(\overrightarrow{\sum e^2}\right)}{rows(e)-2}}$ $\quad se\_e(e)=0.88$

$$R2_{pi.inf}:=1-\frac{\overrightarrow{\sum(e)^2}}{\overrightarrow{\sum(y-mean(y))^2}}$$ $\quad R2_{pi.inf}=0.101$

Ipoteza statistica $\mathbf{H_0:}\ \beta_1=0$ se poate testa folosind statistica t

$t=\frac{b_1}{se\_b1(x,y)}$ urmeaza o distributie t student cu $(n-2)$ grade de libertate. Valuarea *se_b1(x,y)* este abaterea standard a erorii pentru parametrul *b1*

$$se\_b1(x,y):=\frac{se\_e(e)}{\sqrt{\overrightarrow{\sum(x-mean(x))^2}}}$$ $\quad se\_b1(x,y)=0.024$

$$t\left(x,y,\beta_1\right):=\frac{slope(x,y)-\beta_1}{se\_b1(x,y)}$$ $\quad t(x,y,0)=-1.643$

Să incercăm şi o regresie multiplă. Datele folosite sunt producţia industrială pi (variaţie lunară,%), şomajul şomaj (rată lunară,%) şi inflaţia inf (rată lunară,%) pentru perioada ianuarie 2000 - ianuarie 2002. In ordinea dată ele se regăsesc şi in matricea DATE.

Variabila dependenta inflatia, adica :

$y := DATE^{\langle 2 \rangle}$

marimea esantionului: $n := rows(DATE)$

Numarul coloanelor ce reprezinta variabilele indepedente(productia industriala si somajul):

$x\_col := \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $p := rows(x\_col)$ $p = 2$

Matricea X a variabilelor indepedente:

$X := \left\| \begin{array}{l} \text{for } i \in 0..n-1 \\ \quad X_{i,0} \leftarrow 1 \\ \text{for } j \in x\_col \\ \quad X^{\langle j+1 \rangle} \leftarrow DATE^{\langle j \rangle} \\ X \end{array} \right.$

Vectorul parametrilor estimati:

$b := \left(X^T \cdot X\right)^{-1} \cdot \left(X^T \cdot y\right)$

Ecuatia regresiei si erorile reziduale

$ylin := X \cdot b$ $e := y - ylin$

Suma patratelor erorilor si abaterilor:

$SSE := e^T \cdot e$ suma patratelor erorilor

suma patratelor abaterii date de regresie:

$SSR := (ylin - mean(y))^T \cdot (ylin - mean(y))$

suma patratelor abaterii/deviatiei totale

$SST := SSE + SSR$

Gradele de libertate

Cum avem $(p + 1)$ parametrii de estimat avem

$grL\_erori := n - (p + 1)$ $grL\_erori = 23$

grade de libertate asociate cu suma patratelor erorilor. Suma patratelor abaterii/deviatiei date de regresie are

$grL\_regresie := p$ $grL\_regresie = 2$

grade de libertate deoarece in formula sunt p variabile independente .

Suma patratelor abaterii/deviatiei totale are

$grL\_total := n - 1$ $\quad grL\_total = 25$

grade de libertate si $grL\_erori + grL\_regresie = grL\_total$

Pentru a estima dispersia impartim suma patratelor la nr. de grade de libertate. Media abaterilor patratice pentru erorile reziduale

$$MSE := \frac{SSE_0}{grL\_erori}$$

este o estimatie pentru dispersia erorilori σ^2, sau abaterea neexplicata de regresie. (Vom identifica SSE ca fiind primul element din vector.) Alti doi estimatori pentru dispersia care este explicata de modelul regresiei respectiv dispersia totala sunt:

$$MSR := \frac{SSR_0}{grL\_regresie} \qquad MST := \frac{SST_0}{grL\_total}$$

Testul F

In cazul ipotezei nule

H_0: nu avem model de tip regresie multipla

statistica $F := \frac{MSR}{MSE}$ are o distributiie F cu $n1 := grL\_regresie$ si $n2 := grL\_erori$ grade de libertate.

Valuarea p-value a testului este data de :

$p\_val := 1 - pF(F, n1, n2)$ $\quad p\_val = 0.063$

$R2 := \frac{SSR_0}{SST_0}$ si $R2\_ajustat := 1 - \frac{MSE}{MST}$ dau masura potrivirii modelului de regresie liniara.

Analiza Abaterilor

GradeLibertate	SS	MS	F	R2
grL_regresie = 2	SSR = 4.408	MSR = 2.204	F = 3.117	R2 = 0.213
grL_erori = 23	SSE = 16.262	MSE = 0.707	p-value	R2_ajustat = 0.145
grL_total = 25	SST = 20.67	MST = 0.827	p_val = 0.063	

Abaterea explicata de regresia liniara (MSR) este mai mare decit cea datorata erorilor reziduale (MSE). Diferenta este destul de mare (valoarea p-value este suficient de mica) pentru a respinge ipoteza nula.

Corelatia dintre variabilele modelului este data in matricea CORR

$$j := 0..p \qquad k := 0..p \qquad CORR_{j,k} := corr\left(DATE^{<j>}, DATE^{<k>}\right)$$

$$CORR = \begin{matrix} & x1 & x2 & x3 \\ \begin{matrix} \\ \\ \\ \end{matrix} & \begin{bmatrix} 1 & 0.268 & -0.318 \\ 0.268 & 1 & 0.237 \\ -0.318 & 0.237 & 1 \end{bmatrix} & \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} \end{matrix}$$

Cea mai mare valoare a corelatiei dintre variabilele independente este cea intre variabilele (*x1* and *x2*)

$$CORR_{1,0} = 0.268$$

Teste T

$$b = \begin{bmatrix} 0.522 \\ -0.052 \\ 0.196 \end{bmatrix} \qquad \text{unde} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

Forma statisticii t este aceeasi ca si in cazul regresiei simple. In ipoteza nula, $t = \frac{b_1}{se\_b1}$

urmeaza o distributie t student cu (n - 2) grade de libertate . Valuarea *se_b1* este abaterea standard a erorii pentru parametrul *b1*

Matricea Varianta-Covarianta pentru Parametrii Estimati

Fiecare estimatie are o dispersie/varianta dar si o covarianta cu fiecare dintre celelalte estimatii. Acestea se pot reprezenta intr-o matrice Var_Covar_b data de $\sigma^2 \cdot \left(X^T \cdot X\right)^{-1}$

Putem estima varianta necunoscuta σ^2 prin abaterea medie patratica "mean square error: (MSE),

$$Var\_Covar\_b := \left(X^T \cdot X\right)^{-1} \cdot MSE$$

$$Var\_Covar\_b = \begin{matrix} & b_0 & b_1 & b_2 \\ \begin{matrix} \\ \\ \\ \end{matrix} & \begin{bmatrix} 1.251524 & 0.006805 & -0.119521 \\ 0.006805 & 0.000589 & -0.000703 \\ -0.119521 & -0.000703 & 0.01167 \end{bmatrix} & \end{matrix}$$

$Var\_Covar\_b_{0,1} = 6.805 \cdot 10^{-3}$ este covarianta dintre b_0 si b_1. Pe diagonala matricii este varianta/dispersia estimata a coeficientilor.

$Var\_Covar\_b_{1,1} = 5.891 \cdot 10^{-4}$ este o estimatie a variantei/dispersiei lui b_1.

Erorile Standard ale Parametrilor Estimati

Prin definitie vectorul cu erorilor standard ale parametrilor estimati poate fi obtinuta astfel:

$k := 0 .. p$

$se\_b_k := \sqrt{\overrightarrow{Var\_Covar\_b_{k,k}}}$ $\quad se\_b = \begin{bmatrix} 1.119 \\ 0.024 \\ 0.108 \end{bmatrix}$

Statistica T

$t := \frac{\overrightarrow{b}}{se\_b}$ $\quad t = \begin{bmatrix} 0.467 \\ -2.142 \\ 1.81 \end{bmatrix}$

Considerind ipoteza nula,

$H_0: \beta_k = 0$

$k = 0, 1, 2, ..., p$, fiecare test statistic urmeaza o distributie **t** student cu $n - (p + 1) = 23$ grade de libertate, egal cu nr. de grade de libertateal rezidualelor $grL\_erori = 23$

Valuarea corespunzatoare p-value poate fi calculata astfel:

$$p\_t_k := \begin{cases} 2 \cdot \left(1 - pt\left(t_k, grL\_erori\right)\right) & \text{if } t_k > 0 \\ 2 \cdot pt\left(t_k, grL\_erori\right) & \text{otherwise} \end{cases}$$

Tabelul Coeficientilor Regresiei

$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$ $\quad b = \begin{bmatrix} 0.522 \\ -0.052 \\ 0.196 \end{bmatrix}$ $\quad se\_b = \begin{bmatrix} 1.119 \\ 0.024 \\ 0.108 \end{bmatrix}$ $\quad t = \begin{bmatrix} 0.467 \\ -2.142 \\ 1.81 \end{bmatrix}$ $\quad p\_t = \begin{bmatrix} 0.645 \\ 0.043 \\ 0.083 \end{bmatrix}$

Pentru un nivel de semnificatie $\alpha_j := 0.05$

vom respinge ipoteza nula privind neimportanta pentru model a variabilei pentru care conditia este 1

$cond_j := (\alpha_j \geq p\_t_j)$ $\quad cond = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ $\quad \begin{bmatrix} intercept \\ x1 \\ x2 \end{bmatrix}$

Daca conditiile impuse de model sunt indeplinite si nu avem multicoliniaritate, cu un nivel de semnificatie $\alpha = 0.05$ vom pastra in model variabila *x1* si vom exclude *x2* .

Daca incercam un model de tip $\text{inf} = \beta_0 + \beta_1 \cdot \text{pi} + (\beta_1)^2 \cdot \text{somaj} + \varepsilon$

$x1 := \text{DATE}^{<0>}$ $n := \text{rows}(\text{DATE}^{<0>})$

$x2 := \text{DATE}^{<1>}$ $i := 0 .. n - 1$

$y := \text{DATE}^{<2>}$

Modelul este $y = \left[\beta_0 + \beta_1 \cdot x1 + (\beta_1)^2 \cdot x2 + \varepsilon\right]$

este intrinsec nelinear deoarece parametrii apar ca puteri, deci neliniar.

In aceste cazuri intrinsec neliniare putem aplica metoda celor mai mici patrate.

$b_0 := 0$ $b_1 := 0$

Given

$$\sum_i \left(-2 \cdot y_i + 2 \cdot b_0 + 2 \cdot b_1 \cdot x1_i + 2 \cdot b_1^2 \cdot x2_i\right) = 0$$

$$\sum_i \left[2 \cdot \left(y_i - b_0 - b_1 \cdot x1_i - b_1^2 \cdot x2_i\right) \cdot \left(\left(-x1_i - 2 \cdot b_1 \cdot x2_i\right)\right)\right] = 0$$

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} := \text{Find}(b_0, b_1)$$

$b_0 = 2.508$ $b_1 = -0.041$

si calculam $SSY := \overrightarrow{\sum y^2}$ $SSY = 182.67$ cu n grade de libertate $grL\_total := n$

$$SSE := \overrightarrow{\sum \left[y - \left[b_0 + b_1 \cdot x1 + (b_1)^2 \cdot x2\right]\right]^2}$$

$SSE = 18.54$ cu n-p grade de libertate $grL\_erori := n - p$

Pentru variatia explicata de model avem SSR,

$SSR := SSY - SSE$ $SSR = 164.13$ cu p grade de libertate $grL\_regresie := p$

$$MSE := \frac{SSE}{grL\_erori} \qquad MSR := \frac{SSR}{grL\_regresie}$$

Analiza tabelului variatiilor

Grade Libertate	SS	MS
$grL\_regresie = 2$	$SSR = 164.13$	$MSR = 82.065$
$grL\_erori = 24$	$SSE = 18.54$	$MSE = 0.773$
$grL\_total = 26$	$SSY = 182.67$	

Matricea Varianta-Covarianta

Jacobianul:

$$J_{i,0} := -\frac{d}{d\,b_0}\left[b_0 + b_1 \cdot x1_i + (b_1)^2 \cdot x2_i\right]$$

$$J_{i,1} := -\frac{d}{d\,b_1}\left[b_0 + b_1 \cdot x1_i + (b_1)^2 \cdot x2_i\right]$$

si inmultind cu media patratelor erorilor obtinem matricea

$$Var\_Covar := MSE \cdot \left(J^T \cdot J\right)^{-1}$$

$$Var\_Covar = \begin{bmatrix} 0.026 & -2.869 \cdot 10^{-4} \\ -2.869 \cdot 10^{-4} & 5.633 \cdot 10^{-4} \end{bmatrix}$$

Luam apoi radacina patratica din elementele de pe diagonala pentru a obtine erorile standard pentru parametrii estimati

$$j := 0 .. p - 1$$

$$se_j := \sqrt{Var\_Covar_{j,j}} \qquad se = \begin{bmatrix} 0.161 \\ 0.024 \end{bmatrix}$$

ca apoi sa formam intervalele de confidenta pentru b_0,

$$\alpha_t := 0.05$$

$$t := qt\left(1 - \frac{\alpha_t}{2}, grL\_erori\right)$$

$$int\_stinga_{b0} := b_0 - t \cdot se_0 \qquad int\_dreapta_{b0} := b_0 + t \cdot se_0$$

si apoi pentru b_1,

$$int\_stinga_{b1} := b_1 - t \cdot se_1 \qquad int\_dreapta_{b1} := b_1 + t \cdot se_1$$

Avem $1 - \alpha_t = 95\,\%$ siguranta, ca valuarea lui β_0 se afla intre valorile

$$int\_stinga_{b0} = 2.176 \quad si \qquad int\_dreapta_{b0} = 2.84$$

si ca valuarea lui $\beta 1$ se afla intre valorile

$$int\_stinga_{b1} = -0.09 \qquad si \qquad int\_dreapta_{b1} = 7.929 \cdot 10^{-3}$$

Parametrii Estimati	**Erorile Standard**	**Intervalele de incredere**	
$b_0 = 2.508$	$se_0 = 0.161$	$int\_stinga_{b0} = 2.176$	$int\_dreapta_{b0} = 2.84$
$b_1 = -0.041$	$se_1 = 0.024$	$int\_stinga_{b1} = -0.09$	$int\_dreapta_{b1} = 7.929 \cdot 10^{-3}$

Cum intervalul prntru b1 contine valoarea 0 concluzionam ca acest parametru nu este semnificativ pentru model. Deci modelul nu este potrivit.

Bibliografie

*** **MathCad Resource Center**

Robert S Pindyck, Daniel L. Rubinfield **Econometric Models and Econometric Forecasts**, *McGraw-Hill,Inc., International Edition 1991*

Edmond Malinvaud **Methodes statistiques de l'econometrie,** *editura Dunod, 1964*

George Daniel Mateescu **Bazele utilizarii calculatoarelor**, *Editura Donaris 2004*