Analiza datelor de marketing utilizand S.P.S.S.

- <u>analiza predictiva</u> -

Analiza predictiva

- Presupune realizarea de estimari asupra evolutiei viitoare a fenomenelor de marketing, utilizand ca metode de lucru:
    - Analiza seriilor dinamice (*univariata*)
    - Regresia (bivariata sau multivariata)
        - liniara;
        - logistica;
        - hiperbolica;
    - Modelarea.

Criterii de clasificare ale analizei predictive

- Gradul de cuprindere la care se face previziunea:
  - nivel de produs (marca);
  - nivel de grup de produse (linie sau gama);
  - nivel de unitate economica;
  - nivel de ramura de activitate;
  - nivelul economiei nationale (previzune macro-economica);
- Aria geografica inclusa in procesul de previziune:
  - nivel local;
  - nivel regional;
  - nivel national;
  - nivel international.

Criterii de clasificare ale analizei predictive

- Orizondul de previziune poate fi:
  - scurt (o perioada/1 an);
  - mediu (pana la 5 perioade/ani);
  - lung (peste 5 perioade/ani);
- Alte criterii:
  - Precizia rezultatelor (previziuni cantitative si calitative);
  - Tipul de date utilizate;
  - Considerarea influentelor unor factori perturbatori (metode endogene si exogene);

Lanturile Markov

- Metoda lanturilor Markov reprezinta o modalitate de previziune cu utilitate limitata, ce nu presupune nici existenta unei serii cronologice, nici existenta unei asocieri.
- Proprietatea Markov: starea viitoare depinde doar de starea prezenta si de o matrice a probabilitatilor de schimbare a starii (starea viitoare nu depinde de stari trecute) – viitorul este conditional independent de trecut.
- Probabilitatea unei anumite stari de a depinde de starile anterioare:

$$P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1}) = P(s_{ik} \mid s_{ik-1})$$

Lanturile Markov

- Probabilitatea unei stări poate fi calculată cu ajutorul următoarei formule:

$$
\begin{aligned}
P(s_{i1}, s_{i2}, \ldots, s_{ik}) &= P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1})\, P(s_{i1}, s_{i2}, \ldots, s_{ik-1}) \\
&= P(s_{ik} \mid s_{ik-1})\, P(s_{i1}, s_{i2}, \ldots, s_{ik-1}) = \ldots \\
&= P(s_{ik} \mid s_{ik-1})\, P(s_{ik-1} \mid s_{ik-2}) .. P(s_{i2} \mid s_{i1})\, P(s_{i1})
\end{aligned}
$$

- Pentru a defini lanțul Markov trebuie specificate :
  - probabilitatea de tranziție: $a_{ij} = P(s_i \mid s_j)$
  - probabilitatea inițială: $\pi_i = P(s_i)$

Lanturile Markov

- Matricea probabilitatilor de tranzitie este alcatuita pe baza probabilitatile de transformare (schimbare a starii) a fiecarei variabile:
  - Exemplu: utilizarea clasica in marketing – evolutia cotei de piata (matricea probabilitatii de tranzitie este alcatuita pe baza unui indicator de loialitate / tranzitie a respondentilor pentru o anumita marca).
  - Pe piața şampoanelor dermato-cosmetice există trei produse (2007): Selegel, T-gel şi Nizoral, cu cotele de piata:

Selegel	Ducray	Nizoral
25%	35%	40%

Lanturile Markov

- Indicele de loialitate.

Selegel	Ducray	Nizoral
0,85	0,75	0,8

- Probabilitatile de tranzitie (cumparatori care isi vor schimba samponul in luna urmatoare):

Produsul părăsit	Reorietări		
	Selegel	Ducray	Nizoral
Selegel	x	0.10	0.05
Ducray	0.15	x	0.10
Nizoral	0.10	0.10	x

Lanturile Markov

- Matricea probabilitatilor de tranzitie.

0,85	0.10	0.05
0.15	0,75	0.10
0.10	0.10	0,8

- Cotele de piata la t_1:

$$\text{Selegel} = 25 \times 0{,}85 + 35 \times 0{,}10 + 40 * 0{,}05 = 30{,}5$$

Selegel	Ducray	Nizoral
30,5%	32,75%	36,75%

Analiza seriilor dinamice

- Cunoscuta in literatura de specialitate si sub denumirea de analiza seriilor de timp.
- Presupun utilizarea unor date istorice (inregistrari ale evolutiei unui fenomen in timp).
- Reprezinta cea mai facila metoda (logistic si matematic) de realizare a previziunilor.
- Previziunea naiva: in perioada urmatoare variabila investigata isi va pastra nivelul actual:

$$P_{t+1} = Y_t$$

Metoda modificarii procentuale

- Metoda modificării procentuale (MMP) urmăreşte să evalueze schimbarea procentuală a variabilei între perioade succesive de timp.

$$P_{t+1} = t \times MMP_t + Y_0$$

  – unde: MMP_t reprezinta media modificării procentuale pentru primele t perioade, iar Y_0 este valoarea observată din prima perioada a variabilei previzionate.

Metoda modificarii procentuale

- Exemplu: Presupunand un volum al desfacerilor (vanzari) pentru berea Tuborg in primele 6 luni ale anului conform tabelului de mai jos, se vor estima vanzarile din luna iulie.

Luna	Vanzari (hl)
Ianuarie	12000
Februarie	10000
Martie	11000
Aprilie	13000
Mai	14000
Iunie	15000

Metoda modificarii procentuale

- Exemplu: Presupunand un volum al desfacerilor (vanzari) pentru berea Tuborg in primele 6 luni ale anului conform tabelului de mai jos, se vor estima vanzarile din luna iulie.

$$MMP_t = \frac{Y_t - Y_0}{n-1}$$

$$MMP_6 = \frac{15000 - 12000}{6-1} = 600$$

$$Y_{iulie} = 12000 + (7-1) \times 600 = 15600$$

Metoda modificarii procentuale mobile

- Metoda modificării procentuale mobile (MMPM) are un grad mai mare de precizie decat MMP si este utilizata in cazul in care se observa tendinte (trend-uri) in date.
- MMPM presupune calculul prealabil al indicilor care exprimă modificarea procentuală a variabilei de la o perioadă la alta.
- De asemenea, presupune calculul prealabil al mediilor mobile ale schimbarilor procentuale (MPM), dupa formula:

$$MPM_t = \frac{\frac{Y_t - Y_{t-1}}{Y_{t-1}} + \frac{Y_{t-1} - Y_{t-2}}{Y_{t-2}} + \ldots + \frac{Y_2 - Y_1}{Y_1}}{n}$$

Metoda modificarii procentuale mobile

- Metoda modificării procentuale mobile (MMPM) presupune utilizarea formulei de previziune:

$$P_{n+1} = (1 + MMP_n)Y_n$$

- Pentru perioada m care urmeaza celor n perioade observate (date istorice), formula se transforma dupa:

$$P_{n+m} = MMP_n \cdot Y_n \cdot m + Y_n$$

Metoda mediilor mobile

- Metoda mediilor mobile (MM) este utilizata atunci cand se doreste acordarea unei importante (greutati) superioare observatiilor recente dintr-un set de date istorice, fata de cele de la inceputul setului.
- Previziunile se fac asupra unui set de valori ajustate (teoretice), care inlocuiesc termenii initiali ai seriei cronologice, determinate cu ajutorul formului:

$$\hat{Y}_t = \frac{1}{L} \sum_{i=\frac{t-L}{2}}^{\frac{L-1}{2}} Y_t$$

  - presupunea alegerea unui interval de referinta L ($L < n$), la nivelul caruia se vor raporta calculele pentru determinarea mediilor mobile. Se recomanda ca $L < 8$.

Metoda mediilor mobile

- Pentru o serie de aplicatii, se pot utiliza si date “viitoare”, metoda fiind centrata pe o anumita valoare. In acest fel, metoda nu prevede evolutia ulterioara a fenomenului, ci valorile “asteptate”, conform trend-urilor presupuse de valoarile observate.
- Metoda se bazeaza pe propritatea mediei aritmetice de compensare a erorilor, diminuand astfel influenta oscilatiilor periodice. Sirul obtinut reprezinta trendul si reflecta tendinta comuna, generala a seriei cronologice.

Metoda mediilor mobile

- Exemplu: analiza vanzarilor (milioane EURO) lunare ale URBB Bucuresti.

Perioada	1	2	3	4	5	6	7	8	9	10	11	12
Valori observate	5	6	8	7	6,5	7,2	6,8	6,3	6	6,6	7,4	7,8
Valori previzionate (L=5)	-	-	6,5	6,9	7,1	6,8	6,6	6,6	6,6	6,8	-	-

- Metoda de calcul:

$$P_3 = \frac{1}{5}\sum_{i=1}^{5} Y_t = \frac{1}{5}(5+6+8+7+6,5) = 6,5$$

$$P_4 = \frac{1}{5}\sum_{i=2}^{6} Y_t = \frac{1}{5}(6+8+7+6,5+7,2) = 6,9$$

$$P_5 = \frac{1}{5}\sum_{i=3}^{7} Y_t = \frac{1}{5}(8+7+6,5+7,2+6,8) = 7,1$$

Metoda mediilor mobile

- Previziunea se face asupra setului de date ajustat, utilizand metode de analiza a seriilor dinamice la alegere (MMP, MMPM, etc.).
- Media mobila a schimbarilor procentuale (MPM) pentru setul de valori ajustate dupa metoda mediilor mobile este:

$$MPM_t = \frac{\frac{Y_t - Y_{t-1}}{Y_{t-1}} + \frac{Y_{t-1} - Y_{t-2}}{Y_{t-2}} + \ldots + \frac{Y_2 - Y_1}{Y_1}}{n} = 0.06125$$

$$P_{13} = MMP_{10} \cdot \hat{Y}_{10} \cdot 3 + \hat{Y}_{10} = 6.692$$

Metoda nivelarii exponentiale

- Metoda nivelarii exponentiale este mai precisa decat metodele anterioare. La randul ei, creaza posibilitatea ca cele mai recente observatii sa fie luate în calcul cu ponderi mai mari.

$$P_{t+1} = \alpha Y_t + (1 - \alpha)\, P_t$$

- presupunea alegerea unui coeficient de nivelare α (0 < α < 1), valoarea acestuia fiind stabilita fie prin utilizarea mediilor mobile, fie prin incercari, urmata de evaluarea acuratetei seriilor de valori previzionate (suma patratelor valorilor reziduale).

Metoda nivelarii exponentiale

- Exemplu: analiza vanzarilor (milioane EURO) lunare ale URBB Bucuresti. Vom analiza trei coeficienti:
  - α = 0,5;
  - α = 0,33;
  - α = 0,25;

$$P_2 = 0,5 \times 6 + (1 - 0,5) \times 5$$

Perioada	1	2	3	4	5	6	7	8	9	10	11	12
Valori observate	5	6	8	7	6,5	7,2	6,8	6,3	6	6,6	7,4	7,8
Previziune (α=0,5)	5	5,5	6,75	6,9	6,7	6,9	6,9	6,6	6,3	6,4	6,9	7,4
Previziune (α=0,33)	5	5,33	6,22	6,48	6,49	6,73	6,75	6,6	6,4	6,47	6,78	7,12
Previziune (α=0,25)	5	5,25	5,94	6,2	6,28	6,51	6,58	6,51	6,38	6,44	6,68	6,96

Metoda nivelarii exponentiale

- Valorile asteptate pentru perioada urmatoare:
  - 7,6 milioane (α = 0,5);

$$P_{13} = 0,5 \times 7,8 + (1 - 0,5) \times 7,4 = 7,6$$

  - 7,34 milioane (α = 0,33);

$$P_{13} = 0,33 \times 7,8 + (1 - 0,33) \times 7,12 = 7,34$$

  - 7,18 milioane (α = 0,25);

$$P_{13} = 0,25 \times 7,8 + (1 - 0,25) \times 6,96 = 7,18$$

- Pe care o vom alege?

Metoda nivelarii exponentiale

- Metoda nivelarii exponentiale duble (Metoda Brown) este recomandabila atunci cand seria dinamica poseda în configuratia sa o tendinta liniara.
- Necesita doar un minim de 3 valori istorice pentru a fi implementate (insa acuratetea ei este influentata direct de dimensiunea seriei istorice utilizate).
- presupunea utilizarea a doi vectori de nivelare dinamica α_i si β_i ($0 < \alpha_i, \beta_i < 1$).

Metoda nivelarii exponentiale

- Pentru previzionarea unei valori ulterioare k momentului actual (t), se utilizeaza formula:

$$P_{t+k} = \alpha_t + \beta_t P_{k-1}$$

- unde:

$$a_t = 2P_t' - P_t'' \qquad \beta_t = \frac{\alpha}{1-\alpha}\left(P_t' - P_t''\right)$$

- iar

$$P_t' = \alpha X_t + \left(1-\alpha\right)P_{t-1}'$$

$$P_t'' = \alpha\, P_t' + \left(1-\alpha\right)P_{t-1}''$$

Metoda nivelarii exponentiale

- Metoda nivelarii exponentiale cu doi parametrii (Metoda Holt) este mai flexibilitata decat metoda Brown, intrucat permite nivelarea tendintei folosind un parametru diferit de cel al seriei dinamice inițiale.
- Necesita doar un minim de 3 valori istorice pentru a fi implementate (insa acuratetea ei este influentata direct de dimensiunea seriei istorice utilizate).
- presupunea utilizarea a 3 coeficient de nivelare dinamici α, β si γ (0 < α, β, γ < 1).
- Metoda este utilizata pentru a determina trend-ul evolutiei fenomenului, iar pe baza acestuia nivelul ulterior al variabilei previzionate.

Metoda nivelarii exponentiale

- Seriile asociate metodei Holt au forma:

$$P_t = (\alpha + \beta_t) T_t + \varepsilon_t$$

- unde α reprezinta o constanta subunitara asociata nivelului initial al seriei, β este un indice asociat trend-ului seriei, iar ε_t este asociat erorilor (influentelor) aleatorii.
- T_t reprezinta trend-ul (evolutia) asociat seriei de valori istorice observate, calculat dupa formula:

$$T_t = \gamma (P_{t-1} - P_{t-2}) + (1 - \gamma) P_{t-1}$$

Metoda nivelarii exponentiale

- Previziunea valorilor, conform metodei **Holt**, presupune utilizarea formulei:

$$P_t = \alpha Y_t + (1-\alpha)(P_{t-1} + T_t)$$

- In cazul in care in setul de date este inclus si un factor de sezonalitate, se utilizeaza metode nivelarii exponentiale sezoniere a lui Winters.
- Previziunea cu ajutorul acestei metode se bazeaza pe formula:

$$P_{t+m} = (P_t + b_t m) S_{t-L+m}$$

Metoda nivelarii exponentiale

- Sezonalitatea in modelul Winters este estimata cu ajutorul formulei:

$$S_t = \beta \frac{Y_t}{P_t} + (1-\beta) S_{t-1}$$

- unde

$$P_t = \alpha \frac{Y_t}{T_{t-1}} + (1-\alpha)(P_{t-1} + T_{t-1})$$

$$T_t = \gamma (P_t - P_{t-1}) + (1-\gamma) T_{t-1}$$

Alegerea metodei de previziune adecvata

- Selectia modelului de previziune adecvat este realizata prin compararea valorilor reziduale (denumite si variatia neexplicata), dupa formula:

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

- daca metoda utilizata este perfecta, atunci $SS_E = 0$.
- Alternativ, se poata utiliza abaterea medie absoluta (AMA) asociata fiecarei metode de previziune:

$$AMA = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

Alegerea metodei de previziune adecvata

- Exemplu: previziunea vanzarilor pentru a 11-a perioada:

		MMP		Brown		Holt		Winters	
Anul	X_i	Y_i	ε_i	Y_i	ε_i	Y_i	ε_i	Y_i	ε_i
Ian	2	1,8	0,2	2	0	2,3	-0,3	-	-
Feb	2,5	2,3	0,2	2,7	-0,2	2,8	-0,3	2,5	0
Mar	3,2	2,8	0,4	3,3	-0,1	3,4	-0,2	3,1	0,1
Apr	3,0	2,9	0,1	3,1	-0,1	3,2	-0,2	3,1	-0,1
Mai	4,0	3,8	0,2	3,8	0,2	3,8	0,2	3,7	0,3
Iun	4,5	4,6	-0,1	4,6	-0,1	4,4	0,1	4,4	0,1
Iul	5,0	5,2	-0,2	4,8	0,2	4,8	0,2	5,0	0
Aug	4,8	5,0	-0,2	5,3	-0,5	5,0	-0,2	5,1	-0,3
Sep	5,3	5,5	-0,2	5,5	-0,2	5,1	0,2	5,2	0,1
Oct	6,0	5,7	-0,3	5,6	0,4	5,8	0,2	5,5	0,5

Alegerea metodei de previziune adecvata

- Suma patratelor valorilor reziduale, respectiv abaterea medie absoluta:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad AMA = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

	MMP	Brown	Holt	Winters
SS_E	0,51	0,6	0,47	0,47
AMA	2,1	0,2	0,21	0,17

Modele autoregresive (AR)

- Modelele autoregresive reprezinta o varianta univariata a regresiei liniare, in care valoarea curenta este estimata utilizand una sau mai multe valori anterioare ale seriei (serii cronologice).
- Modelul AR: $\hat{Y}_t = \delta + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \varepsilon_t$
- unde p reprezinta ordinul de autoregresie (nivelarea exponentiala reprezinta un model AR de ordin 1), δ este un indice asociat trend-ului seriei, iar ε_t este asociat erorilor (influentelor) aleatorii.

$$\delta = \left(1 - \sum_{i=1}^{p} \alpha_i\right) \bar{Y}$$

Modele autoregresive (AR)

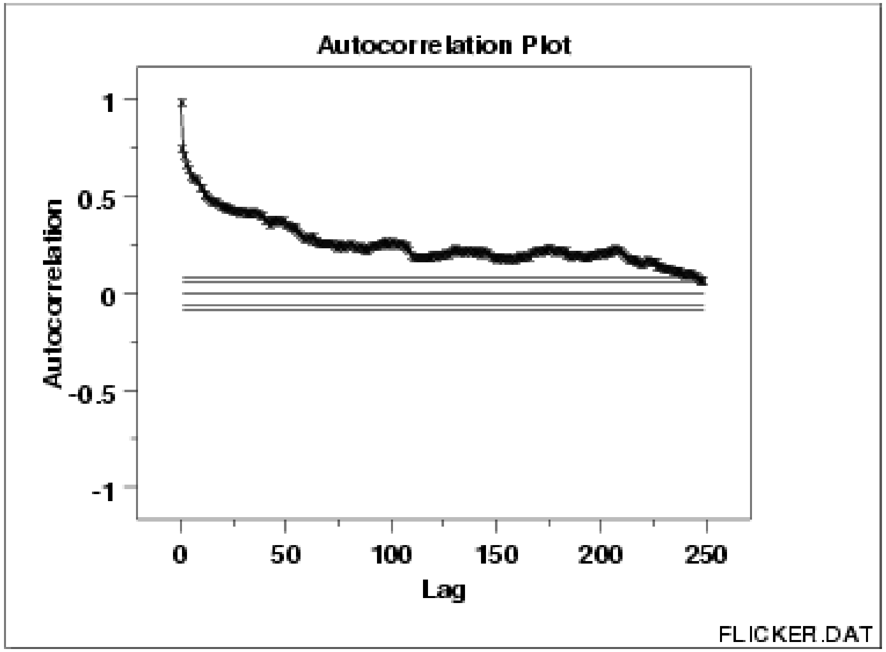
- Box & Jenkins au demonstrat ca una dintre cele mai eficiente modalitate de rezolvare a modelelor autoregresive este prin utilizarea mediilor mobile (Moving Averages – MA).
- Variantele metodei Box-Jenkins:
  - ARMA – utilizat pentru serii stationare (*serii cu proprietatea ca media si variatia nu se modifica semnificativ in timp – practic, o serie de tip Brown, in care nu exista trend si sezonalitate*).
  - ARIMA – utilizat pentru serii dinamice ("I" vine de la Integrate).

Modele autoregresive (AR)

- Metoda Box-Jenkins presupune trecerea prin 3 faze pentru determinarea modelului utilizat in previziune:
  1. Identificarea modelului
  2. Estimarea parametrilor modelului
  3. Validarea modelului
- In general, pentru realizarea unei autoregresii eficiente, sunt recomandate serii cronologice lungi – unii autori recomanda minim 50 de observatii, alti chiar 100.

Modele autoregresive (AR)

- Identificarea modelului:
  - Dinamicitatea unei serii (modelul ARMA sau ARIMA) este determinata utilizand un grafic de autocorelatie, care va prezenta sezonalitate in cazul in care graficul este continuu



Modele autoregresive (AR)

- Identificarea modelului:
  - Graficul de autocorelatie reprezinta pe abcisa trecerea timpului, iar pe ordonata coeficientul de auto-corelatie corespunzator, calculat dupa formula:

$$R_h = \frac{\frac{1}{n}\sum_{i=1}^{N-h}(Y_t - \bar{Y})\,(Y_{t+h} - \bar{Y})}{\sigma^2}$$

  - Liniile (valorile) de demarcatie pentru autocorelatie sunt calculate dupa formula (α corespunde probabilitatii de garantare a rezultatelor):

$$\pm \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

Modele autoregresive (AR)

- Identificarea modelului:
  - Modelul ARMA (fara sezonalitate si trend):

$$\left(1-\sum_{i=1}^{p}\alpha_i L^i\right) Y_t = \left(1+\sum_{i=1}^{q}\beta_i L^i\right)\varepsilon_t$$

  - Modelul ARIMA (serii dinamice):

$$\left(1-\sum_{i=1}^{p}\alpha_i L^i\right)(1-L)^d Y_t = \left(1+\sum_{i=1}^{q}\beta_i L^i\right)\varepsilon_t$$

Modele autoregresive (AR)

- Identificarea modelului:
  - estimarea parametrilor α_i si β_i - in intervalul [-1;1] se realizeaza prin aproximare (recomandabil cu un program statistic, gen SPSS);
  - L_i reprezinta vectorul primilor i parametrii estimati pentru o serie cronologica simpla sau care include sezonalitate (operatorul de lag).
- Estimarea parametrilor modelului:
  - parametrii p si q sunt estimati cu ajutorul graficului de autocorelatie (valoarea maxima a lui α (probabilitatea de garantare a rezultatelor) pentru care coeficientii de autocorelatie nu depasesc valoarea-prag).
  - parametrii α_i sunt estimati prin aproximare, folosind metoda celor mai mici patrate (recomandabil cu un program statistic, gen SPSS);

Modele autoregresive (AR)

- Validarea parametrilor modelului:
  - Se realizeaza prin testarea ipotezei nule ca valorile reziduale sunt independente, vectorul acestora avand o medie si o varianta nediferite semnificativ statistic in timp. In cazul in care parametrii nu sunt validati, trebuie revenit la pasul 1.
  - Valoarea testul Student asociat parametrilor modelului este:

$$z_c = \frac{\alpha_i}{s_{\alpha_i}}$$

    - $-Z_t \leq Z_c \leq Z_t$: se accepta ipoteza nula (parametrul NU este valid);
    - altfel, se accepta ipoteza alternativa (parametrul este valid);

Modele autoregresive (AR)

- Exemplu: previziunea vanzarilor pentru a 11-a perioada:

Anul	Y_i (Vanzari mil. $)
Ian	10
Feb	12
Mar	11
Apr	14
Mai	14,5
Iun	15
Iul	16
Aug	18,5
Sep	19
Oct	20

Modele autoregresive (AR)

$$\hat{Y}_t = \delta + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \varepsilon_t$$

- Valoarea coeficientilor de grad 3, estimata de catre SPSS:
  - δ = -0,934
  - α_1 = 0,534 α_2 = - 0,398 α_3 = 1,062
- Ecuatia de autoregresie devine astfel:

$$\hat{Y}_t = -0{,}934 + 0{,}534 Y_{t-1} - 0{,}398 Y_{t-2} + 1{,}062 Y_{t-3}$$

Modele autoregresive (AR)

- Pentru perioada 11 vom avea:

$$\hat{Y}_{11} = -0{,}934 + 0{,}534 \times 20 - 0{,}398 \times 19 + 1{,}062 \times 18{,}5 = 21{,}8$$

- Testarea semnificatiei parametrilor:

$$z_c = \frac{\alpha_3}{s_{\alpha_3}} = \frac{1{,}062}{0{,}333} = 3{,}218$$

  – pentru $\alpha=0{,}05$ $z_t=1{,}96$ => $z_c > z_t$ => ipoteza alternativa va fi acceptata (parametrul este valid)

Modele autoregresive (AR)

- Testarea semnificatiei parametrilor:

$$z_c = \frac{\alpha_2}{s_{\alpha_2}} = \frac{-0{,}398}{0{,}396} = -1{,}005 \quad z_c = \frac{\alpha_1}{s_{\alpha_1}} = \frac{-0{,}534}{0{,}317} = 1{,}684$$

- pentru $\alpha=0{,}05$ $z_t=1{,}96$ =>

  $-z_t$ (-0,96) $\leq z_c$ (-1,005) $\leq z_t$ (1,96) => ipoteza nula va fi acceptata (parametrul NU este valid)

$$\hat{Y}_t = -0{,}934 + 1{,}062 Y_{t-3}$$

$$\hat{Y}_{11} = -0{,}934 + 1{,}062 \times 18{,}5 = 18{,}7$$

Analiza autocorelatiei

- Testul Durbin-Watson necesita calculul parametrului d, dupa formula:

$$d = \frac{\sum_{t-2}^{T} (\hat{U}_t - \hat{U}_{t-1})^2}{\sum_{t=1}^{T} \hat{U}_t^2}$$

  - Daca $d<d_L$ sau $d>d_T$, atunci este acceptata ipoteza nula (d_L si d_T sunt luate din tabelele asociate testului Durbin-Watson).

- Testul Geary este de natura neparametrica si are ca punct de plecare calculul numarului schimbarilor de semn in seria valorilor reziduale δ.
  - Daca $\delta_{min} < \delta < \delta_{max}$ (tabelate), atunci ipoteza nula este acceptata.

Regresia

– Regresia reprezinta o clasa semnificativa de metode de previziune, in care valoarea unei variabile (denumita dependenta) este previzionata folosind valorile altor variabile (independente), de ale carei valori depinde.

– Dependenta variabilei previzionate trebuie demonstrata, utilizand un coeficient de corelatie (corelatia trebuie sa fie cel putin medie, dar se recomanda utilizarea corelatiilor puternice sau foarte puternice).

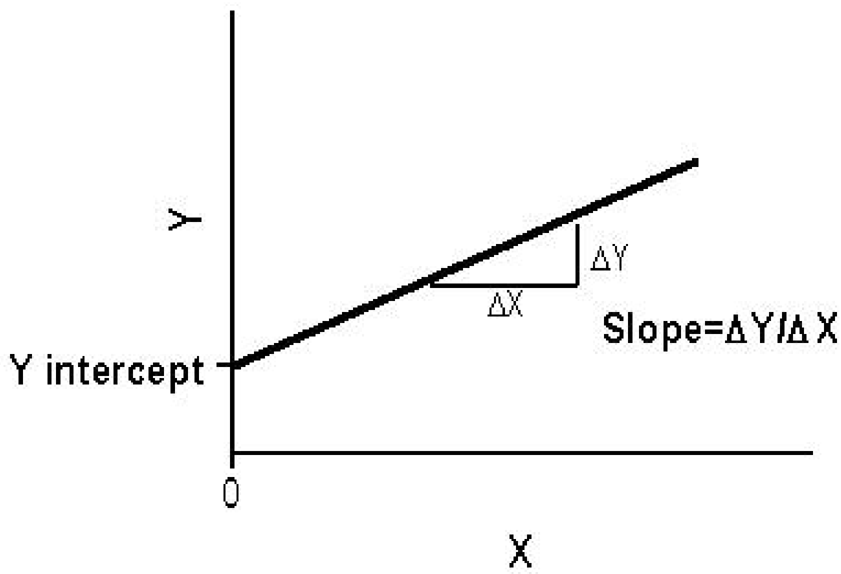
Regresia

- Formele regresiei:
  - in functie de <u>numarul de variabile</u> utilizate:
    - bivariata (o singura variabila independenta);
    - multivariata (doua sau mai multe variabile independente);
  - in functie de <u>forma relatiei dintre variabile</u> (identificata cu ajutorul analizei grafice):
    - liniara;
    - logistica;
    - polinomiala;
    - trigonometrica;

Regresia liniara

- Regresia liniara bivariata:

$$y = a + bx$$



Regresia liniara

- Parametrii regresiei (metoda celor mai mici patrate):
  - panta (b):

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

  - termenul liber (a):

$$a = \bar{y} - b\bar{x}$$

Metoda regresiei multiple

- Permite analiza relatiei liniare dintre o variabila dependenta si una sau mai multe variabile indepentende
- Obiectiv: explicarea si previziunea variatiei variabilei dependente in functie de covarianta ei cu variabilele independente.

$$\hat{Y} = \alpha + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_i X_i + \ldots + \hat{\beta}_n X_n$$

- Parametrii β sunt estimati utilizand metoda celor mai mici patrate (un model cu n variabile va avea nevoie de n perechi de date “istorice” pentru scrierea unui sistem de n ecuatii).

Exemplu: cererea de bunuri/servicii (dependenta) in functie de factori determinanti (venituri, cifra de afaceri, pret, etc.)

Metoda regresiei multiple

- Metoda celor mai mici patrate pentru o regresie liniara de gradul 2:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_{i1} \times y_i)\sum_{i=1}^{n} x_{2i}^2 - \sum_{i=1}^{n}(x_{2i}y_i)\sum_{i=1}^{n}(x_{i1}x_{i2})}{\sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2 - \left(\sum_{i=1}^{n} x_{i1}x_{i2}\right)^2}$$

$$\beta_2 = \frac{\sum_{i=1}^{n}(x_{i2}y_i)\sum_{i=1}^{n} x_{i1}^2 - \sum_{i=1}^{n}(x_{i1}y_i)\sum_{i=1}^{n}(x_{i1}x_{i2})}{\sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2 - \left(\sum_{i=1}^{n} x_{i1}x_{i2}\right)^2}$$

$$\alpha = \bar{y} - \beta_1\bar{x}_1 - \beta_2\bar{x}_2$$

Metoda regresiei multiple

- Estimarea semnificatiei statistice a parametrilor este utilizata pentru a se verifica faptul ca variatia variabilei dependente nu este datorata intamplari (evenimentelor aleatoare), ci este rezultatul variatiei uneia sau mai multor variabile independente.
- Realizata cu ajutorul testului Student, in care numarul de grade de libertate al valorii teoretice (tabelate) se determina cu conform:

Nivelul de semnificatie = (1-nivelul de confidenta)/2

Metoda regresiei multiple

- Testarea semnificatiei (reprezentativitatii) parametrilor de regresie:

$$t_c = \frac{\beta_i}{s_{\beta_i}} \qquad \beta_j \pm s_{\hat{\beta}_j} \times t_{T,j}$$

- Eroarea standard a unui parametru estimat arata cu cat poate sa varieze acesta in jurul valorii sale ca urmare a erorii aleatoare.

Metoda regresiei multiple

- Testul F este utilizat pentru a determina semnificatia (reprezentativitatea) variatiei variabilei dependente explicata de variatia variabilelor independente considerate.
- Utilizeaza formula:

$$F_c = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})(n - k - 1)}{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 (k - 1)}$$

Metoda regresiei multiple

- ***Coeficientul (raportul) de corelație multiplă R*** reprezinta gradul in care variabilele independente, per ansamblu, explica variatia variabilei dependente .
- Utilizeaza formula:

$$R_{y,x_1,x_2,\ldots,x_k} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}$$

Metoda regresiei multiple

- Pentru a putea caracteriza proporția variației variabilei dependente datorată variației setului de variabile variabile independente ale modelului se calculează coeficientul de determinare multiplă R^2 (pătratul raportului de corelație multiplă), care arată proporția din variația totală a variabilei Y care este explicată de variabilele independente X_1, X_2, ...X_k.

- În afara coeficienților de corelație multiplă, în analiza corelației dintre variabile se mai pot calcula şi coeficienții de corelație parțială, ce caracterizează intensitatea legăturii dintre două variabile, în ipoteza că celelalte variabile rămân constante

Metoda regresiei multiple

- Exemplu: Estimarea nivelului vanzarilor de telefoane mobile plecand de la suprafata comerciala a magazinului si numarul de asistenti de vanzare.

Vânzări (bucăți)	Număr vânzători (persoane)	Suprafața comercială
22	7	98
20	5	90
23	8	110
26	9	130
30	12	140
32	15	145
45	22	156
50	25	160
52	32	164
60	40	175

Metoda regresiei multiple

- Sistemul de 3 ecuații simultane cu 3 necunoscute, pentru determinarea estimatorilor α, β_1 şi β_2 este.

$$\begin{cases} n\alpha + \beta_{1i}\sum x_{1i} + \beta_2 \sum x_{2i} = \sum y_i \\ \alpha \sum x_{1i} + \beta_1 \sum x^2{}_{1i} + \beta_2 \sum x_{1i}x_{2i} = \sum x_{1i}y_i \\ \beta \sum x_{2i} + \beta_1 \sum x_{1i}x_{2i} + \beta_2 \sum x^2{}_{2i} = \sum x_{2i}y_i \end{cases}$$

$$\begin{cases} 10\alpha + 175\beta_1 + 1368\beta_2 = 360 \\ 175\alpha_1 + 4321\beta_1 + 26721\beta_2 = 7816 \\ 1368\alpha_1 + 2672\beta_1 + 194786\beta_2 = 52754 \end{cases}$$

Metoda regresiei multiple

- Dupa rezolvarea ecuatiei vom obtine:
  - β_1 = 0,974543752;
  - β_2 = 0,104112437;
  - α = 4,702902918;

$$\hat{Y} = 4,703 + 0,97X_{1i} + 0,104X_{2i}$$

- Coeficientul de corelatie multipla este:

$$R_{y,x_1,x_2,\ldots,x_k} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = 0,989085$$

Metoda regresiei multiple

- Valorile reziduale:

Y_i	$\hat{y}$	y_i	$\varepsilon = y_i - \hat{y}$	$(y_i - \hat{y})^2$
22	22,92209467	22	-0,922094675	0,850258589
20	18,15286921	20	1,847130787	3,411892145
23	23,49930977	23	-0,499309769	0,249310245
26	26,96671515	26	-0,966715154	0,934538188
30	31,04921181	30	-1,04921181	1,100845422
32	34,49973652	32	-2,499736517	6,248682653
50	45,79082822	50	4,209171778	17,71712706
52	52,87302888	52	-0,873028881	0,762179427
60	61,77950786	60	-1,779507855	3,166648206
				40,85910144

Metoda regresiei multiple

- Validitatea valorilor previzionate:

$$F_c = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})(n-k-1)}{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 (k-1)} = 157{,}7125516$$

- Valoarea tabelata a lui F pentru o probabilitate de garantare a rezultatelor de 95% si 52 de grade de libertate: 3,23 => F_c =157,71 > F_t =3,23 => se accepta ipoteza alternative (valoarea coeficientului de corelatie multipla este semnificativ diferita de zero), deci regresia este valida.

Analiza multicoliniaritatii

- Coliniaritatea reprezinta relatia liniara dintre doua variabile independente ale unui model.
- Prezenta sa poate duce la distorsiuni serioase ale parametrilor modelului.
- Sugerata de prezenta erorilor standard mari sau de sensitivitatea exagerata a parametrilor.
- Evidentiata utilizandu-se cele trei teste Farrar si Glauber.

Primul test Farrar si Glauber

– Se bazeaza pe compararea matricei de corelatie a modelului cu matricea unitate, cu ajutorul testului χ^2

$$\chi_c^2 = -\left[n - 1 - \frac{1}{6}\left(2(m-1) + 5\right)\right]\ln\det[Z^T Z]$$

- Valoarea teoretica a lui χ^2 se regaseste in tabelele statistice ale repartitiei χ^2, considerandu-se 1/2(m-1)(m-2) grade de libertate.
- Daca χ^2 > χ^2, atunci se concluzioneaza ca exista multicoliniaritate la nivelul modelului (regresiei) analizate.

Al doilea test Farrar si Glauber

- Permite identificarea variabilelor cel mai afectate de coliniaritate
- Se bazeaza pe compararea matricei de corelatie a modelului cu matricea unitate, cu ajutorul testului Fisher.

$$F_c = (r^{ii} - 1)\frac{(n - (m - 1))}{m - 2}$$

- Valoarea teoretica a lui F se regaseste in tabelele statistice ale repartitiei Fisher, considerandu-se n-m+1 si m-2 grade de libertate.

Daca $F_c > F_t$, atunci se concluzioneaza ca ipoteza ortogonalitatii intre variabilele independente nu este acceptata.

Al treilea test Farrar si Glauber

- Permite stabilirea semnificatiei statistice a coeficientilor de corelatie
- Coeficientii de corelatie partiala intre X_i si X_j se determina pe baza formului:

$$r_{ij} = \frac{-r^{ij}}{\sqrt{r^{ii}} - \sqrt{r^{jj}}}$$

- Apoi se calculeaza valoarea testului Student dupa formula:

$$t_{ij} = \frac{r_{ij} \times \sqrt{n - (m-1)}}{\sqrt{(1 - r_{ij}^2)}}$$

Daca $t_{ij} > t_t$, atunci se concluzioneaza ca ipoteza nula este respinsa.

Analiza erorii medii patratice a valorilor reziduale

- Masura sintetica a acuratetii modelului si o metoda de evidentiere a erorilor de previziune.

$$\frac{1}{T}\sum_{t=1}^{T}(P_t - A_t)^2 = (\overline{P} - \overline{A}) + (S_P - S_A)^2 + 2(1-r)S_P S_A$$

  - $(P\text{-}A)^2$ indica tendinta medie a modelului de a supraestima sau subestima valorile reale.
  - $(S_P\text{-}S_A)^2$ indica sensitivitatea modelului la modificarea valorilor independente.
  - $2(1\text{-}r)S_PS_A$ indica marimea erorii datorate lipsei corelatiei perfecte dintre valorile previzionate si cele actuale.