2.3. Modele ARIMA şi construirea lor prin metodologia BOX-JENKINS

1. Cum modelăm o serie staţioară, adică, ce fel de model de regresie putem folosi pentru a descrie comportamentul unei serii staţionare?
2. Cum folosim modelul ajustat pentru prognoză?

Modelele de tip ARMA pot să aproximeze cele mai multe procese staţionare.
Pentru că în economie sunt puţine serii de timp staţionare, sunt necesare modele care să fie capabile să reproducă comportamentul nestaţionar. Astfel de modele sunt modelele de tip ARIMA (modele autoregresive integrate de medie mobilă), care se obţin prin presupunerea că o serie cronologică poate fi reprezentată, după diferenţiere, printr-un model de tip ARMA.
Multe serii de timp din economie sunt nestaţionare, dar sunt integrate. Dacă o serie de timp este $I(d)$, după diferenţierea seriei de d ori se obţine o serie staţionară, adică $I(0)$. Aplicând un model ARMA acestei serii $I(0)$, spunem că seria de timp originală este un proces de medie mobilă integrat, autoregresiv, notat ARIMA(p,d,q). Se ştie deja că p indică numărul de termeni autoregresivi, q indică numărul de termeni de medie mobilă, iar d arată de câte ori a fost diferenţiată seria pentru a deveni staţionară.
Metodologia Box-Jenkins are ca obiectiv identificarea şi estimarea unui model statistic care poate fi considerat că a generat datele de selecţie. Acest model estimat va fi utilizat pentru prognoze şi este necesară presupunerea că are caracteristici constante în timp. Selectarea modelului ARIMA potrivit pentru datele observate se realizează printr-o procedură iterativă (metodologia Box-Jenkins), bazată pe mai multe etape.
EtapaI. **Identificarea**.
Obiectivul identificării este de a selecta o subclasă a unei familii de modele ARIMA, potrivite pentru a reprezenta o serie de timp, adică un proces economic evolutiv.
În etapa de identificare sunt analizate datele observate şi toate informaţiile disponibile care sunt capabile să sugereze o subclasă de modele ce pot descrie modul de generare a datelor.
Elementele de teorie economică pot sugera tipul de model potrivit pentru un anumit proces economic. Astfel, **procesele economice inerţiale** (**consumul, inflaţia, plata impozitelor**) sau cele dependente de propriile realizări anterioare (**exportul, productivitatea, investiţiile, producţia din sectorul zootehnic, calitatea producţiei**), se recomandă a fi reprezentate de un model AR, ARI, ARMA sau ARIMA.
Procesele economice care pot fi relativ uşor influenţate de şocuri exterioare, dar care permit măsuri compensatorii în timp (**aprovizionarea, producţia, cotaţia la bursă**) se recomandă a fi reprezentate de un model MA sau care include o parte MA.
Identificarea înseamnă determinarea tipului de model şi a ordinului modelului ce urmează a fi folosit pentru a reproduce caracteristicile dinamice ale datelor. Pentru a determina cea mai bună specificare sunt folosite reprezentări grafice ale datelor şi ale funcţiilor ACF şi PACF. Se analizează funcţiile de autocorelaţie şi autocorelaţie parţială.
Mai întâi se examinează seria de analizat, pentru a vedea dacă ar fi putut fi generată de un proces staţionar. De regulă, seriile de timp din economie nu sunt staţioanare. O astfel de serie poate avea următoarele caracteristici:
1.Seria prezintă o medie care nu este constantă în timp, dar variaţiile sunt constante de la o perioadă la alta; urmează, de regulă, o traiectorie liniară, cu panta pozitivă sau negativă. Se spune că este o serie nestaţionară de tip omogen.
2. Seria prezintă variaţii care nu sunt constante de la o perioadă la alta. Se spune că este o serie nestaţionară de tip neomogen.

Folosim reprezentarea grafică a seriei pentru a vedea dacă are varianță staționară. Dacă dispersia datelor variază cu timpul, seria este nestaționară. În acest caz se poate aplica datelor tansformarea Box-Cox.
Urmează analiza staționarității în medie. Acest lucru se poate face examinând graficul valorilor seriei, corelogramele de selecție sau prin aplicarea unui test pentru o rădăcină egală cu 1, precum testul Dickey-Fuller. Autocorelațiile de selecție sunt estimații consistente ale coeficienților populației corespunzători, aşa încât corelograma de selecție a u nui proces staționar tinde la zero pentru lungimi de lag moderate. O serie nestaționară are o ACF care nu descreste. O serie integrată are SACF ce descreşte foarte încet, iar SPACF scade foarte repede, la lagul $k = 2$ şi prima valoare este foarte apropiată de 1.
Dacă seria arată un model de nestaționaritate, se staționarizează seria prin aplicarea unor transformări potrivite. De exemplu, se logaritmează valorile inițiale ale seriei, apoi se diferențiază seria până la obținerea staționarității. Luăm seria diferențelor de ordinul întâi şi analizăm dacă este staționară într-un mod similar. Metodele grafice pot fi însoțite de teste de rădăcină unitară. Acest proces de a efectua diferențele va continua până la obținerea unei serii staționare, z_t.
După obținerea unei serii staționare z_t, se compară SACF şi SPACF ale lui z_t cu diferite modele teoretice ARMA, iar această comparație poate sugera câteva modele plauzibile.
În cazul unui model AR(p) ACF este infinită iar PACF este finită.
Coeficienții de autocorelație descresc spre 0 geometric sau în alternanță, pe măsură ce creşte lagul k.
Coeficienții de autocorelație parțială descresc brusc la 0 după un anumit număr de decalaje, astfel încât pentru $k > p$, $\phi_{kk} \approx 0$.
Rezultă că, dacă coeficienții $\hat{\rho}_k$ scad în progresie geometrică şi coeficienții $\hat{\phi}_{kk}$ scad brusc astfel încât $\hat{\phi}_{k,k} \approx 0$, pentru $k > p$, se recomandă un model AR(p). Deci p este cea mai mare valoare a lui k, pentru care coeficienții de autocorelație parțială estimați, $\hat{\phi}_{kk}$, se află în afara intervalului $(\pm 1{,}96/\sqrt{n})$.
În cazul unui model MA(q), PACF este infinită iar ACF este finită.
Coeficienții de autocorelație scad brusc la 0 astfel încât pentru $k > q$, $\rho_k \approx 0$.
Coeficienții de autocorelație parțială descresc spre 0 exponențial sau cu fluctuații.
Rezultă că, dacă coeficienții $\hat{\rho}_k$ scad brusc astfel încât $\hat{\rho}_k \approx 0$ pentru $k > q$ şi coeficienții $\hat{\phi}_{kk}$ scad în progresie geometrică, se recomandă un model MA(q). Deci q este cea mai mare valoare a lui k, pentru care coeficienții de autocorelație estimați, $\hat{\rho}_k$, se află în afara intervalului $(\pm 1{,}96/\sqrt{n})$.
Pentru un model ARMA(p,q), atât coeficienții de autocorelație cât şi coeficienții de autocorelație parțială descresc spre 0 exponențial sau geometric. Ordinul *p*, al părții autoregresive din model, este dat de coeficientul de autocorelație parțială care prezintă un nivel semnificativ, iar ordinul *q*, al părții de medie mobilă din model, este dat de coeficientul de autocorelație care prezintă un nivel semnificativ. La acest moment, folosirea unui model de tip ARMA(p,q) este doar la stadiul de intenție.
Observație: Scăderea coeficienților de autocorelație trebuie înteleasă în valoare absolută.
Descreştera coeficienților de AC sau PAC poate fi întreruptă de valori mai mari (în valoare absolută) în cazul în care datele prezintă sezonalitate (de ex., în cazul datelor trimestriale, se pot întâlni valori mari ale coeficienților $\hat{\rho}_4$ sau $\hat{\phi}_{44}$, comparativ cu valorile coeficienților $\hat{\rho}_3$, $\hat{\rho}_5$, $\hat{\rho}_6$.

Se recomandă selectarea mai multor variante de modele, printre care şi o variantă ARMA, pentru a fi estimate şi verificate.
- Dacă există o valoare a decalajului, q, după care valoarea funcţiei de autocorelaţie scade brusc spre 0 rezultă că, pentru prelucrarea seriei, este indicată folosirea unui proces MA(q) pur sau a unui proces ce conţine o componentă MA(q).
- Dacă valoarea funcţiei de autocorelaţie parţială $\hat{\rho}_{kk} = \hat{\phi}_{kk}$, scade brusc spre 0 după o valoare a decalajului egală cu p, rezultă că, pentru prelucrarea seriei, este indicată folosirea unui proces AR(p) pur sau a unui proces ce conţine o componentă AR(p). Se consideră că valoarea lui $\hat{\rho}_{kk} = \hat{\phi}_{kk}$ diferă semnificativ faţă de zero dacă nu este conţinută în intervalul $\left(-t_{\frac{\alpha}{2};n-2} \cdot \frac{1}{\sqrt{n}}, +t_{\frac{\alpha}{2};n-2} \cdot \frac{1}{\sqrt{n}}\right)$.

Etapa II. **Estimarea parametrilor modelului identificat**.

Estimarea parametrilor unui model de tip AR(p) se face prin MCMMP, minimizând suma pătratelor reziduurilor sau prin ecuaţiile Yule-Walker de selecţie.
Dacă în model există termeni de tip MA, minimizarea sumei pătratelor reziduurilor sau maximizarea funcţiei de verosimilitate necesită metode de estimare neliniare (Newton, Davidson-Fletcher-Powell).
Parametrii unui model ARIMA(p,d,q), se pot estima consistent prin MCMMP sau prin metoda verosimilităţii maxime. Ambele procedee de estimare sunt bazate pe calculul inovaţiilor ε_t din valorile variabilei staţionare.
Estimarea diverselor variante de modele, care au fost sugerate de corelogramă în etapa de identificare, va fi realizată cu ajutorul unui program ARMA dintr-un pachet de programe (EViews, SAS, SPSS). Vom atribui valori pentru p şi q în diverse combinaţii şi vom obţine estimaţii ale coeficienţilor modelelor corespunzătoare.

Verificarea corectitudinii modelului identificat.

După ce modelele ARMA selectate au fost estimate, dorim să ne asigurăm că sunt potrivite. Pentru aceasta se analizează atât reziduurile cât şi parametrii pentru fiecare model.

Testarea reziduurilor

Practica standard este de a reprezenta grafic reziduurile modelului estimat.
Dacă modelul estimat este adecvat, reziduurile ar trebui să fie aproximativ zgomot alb. Deci, reziduurile obţinute pentru modelul estimat sunt analizate pentru a stabili dacă provin dintr-un proces de zgomot alb. În caz afirmativ, tipul de model propus este o bună aproximaţie pentru procesul stochastic de bază. În caz contrar, procesul este reluat, metoda fiind una iterativă.
Pentru a controla dacă reziduurile au media zero şi sunt necorelate folosim atât graficul, cât şi ACF şi PACF ale reziduurilor. ACF şi PACF teoretice ale procesului WN iau valoarea zero pentru lagurile $k \neq 0$, aşa că, dacă modelul este potrivit, cei mai mulţi din coeficienţii funcţiilor SACF şi SPACF ar trebui să fie apropiaţi de zero. În practică, cerem ca 95% din aceşti coeficienţi să se afle în interiorul unor limite de nesemnificaţie. Mai mult, statistica Ljung-Box ar trebui să aibe valori mici, deoarece corespunde la variabile necorelate.
Un model care conduce la obţinerea unor coeficienţi de autocorelaţie a reziduurilor semnificativi sau la o statistică Q semnificativă la nivelul de semnificaţie de 10%, trebuie eliminat.
Dacă se observă că varianţa reziduurilor este crescătoare, rezultă că este potrivită o transformare logaritmică. Este foarte important ca reziduurile obţinute prin estimarea unui model să fie serial necorelate. Deci, orice model care nu produce reziduuri aleatoare ar trebui să nu fie luat în considerare, ci eliminat.

Testarea parametrilor

Se pot aplica testele de semnificaţie clasice t şi F

Este important, de asemenea, să fie satisfăcute condiţiile de staţionaritate şi inversabilitate. Dacă factorizăm polinoamele AR şi MA şi una din aceste rădăcini este apropiată de 1, aceasta este un semn de LACK a staţionarităţii sau inversabilităţii.

Departajarea între mai multe modele

Presupunând că sunt estimaţi parametrii pentru mai multe modele identificate ca fiind ARIMA(p_i,d,q_i), se pot utiliza două tipuri de indicatori: ai calităţii reziduului sau ai teoriei informaţiei.

a) Indicatori ai calităţii reziduului

Dintre modele estimate se alege modelul cu cea mai mică varianţă a reziduurilor, cu cea mai mare valoare a lui R^2 şi cu cea mai mare valoare a statisticii F.

b) Indicatori ai teoriei informaţiei

Pentru fiecare din modelele estimate se calculează indicatorii AIC(criteriul Akaike) şi BIC (criteriul Schwarz sau bayesian): $\mathrm{AIC} = \log \hat{\sigma}_\varepsilon^2 + 2\frac{p+q}{n}$, $\mathrm{BIC} = \log \hat{\sigma}_\varepsilon^2 + (p+q)\frac{\log n}{n}$.

Se alege modelul cu cea mai mică valoare a indicatorului.

Etapa III. **Efectuarea de prognoze cu ajutorul modelelor ARIMA**.

Obiectivul principal urmărit prin construirea unui model ARMA este de a predicta sau prognoza evoluţia viitoare a unei variabile economice. Modelul selectat, pentru care s-a verificat corectitudinea, poate fi utilizat pentru prognoză. Modelele ARIMA sunt destul de apreciate pentru succesul în a face predicţii. Prognozele obţinute prin aceste modele sunt considerate mai bune, mai ales pe termen scurt, decât cele obţinute prin modelarea econometrică tradiţională, aplicată în cazul seriilor cronologice.

Prognozarea unei serii de timp constă în încercarea de a previziona valori viitoare ale seriei, ţinând seama de valorile din trecut ale seriei şi de valorile din trecut ale unui termen eroare. Se pot efectua **prognoze punctuale** (se previzionază o singură valoare a variabilei) sau **prognoze pe interval** (se determină un domeniu de valori în care anticipăm că se va afla valoarea viitoare a variabilei, pentru un nivel de semnificaţie dat).

Ipoteza cheie: Stationaritatea seriei

- regula de bază care guvernează nu se schimbă
- comportamentul viitor poate fi dedus pe baza trecutului
- avem suficiente observaţii din trecut pentru a înţelege procesul de bază.

Considerăm că y_t este un proces stochastic staţionar cu media zero. Putem presupune că suntem la momentul n şi suntem interesaţi de a prognoza sau previziona valorile acestui proces pentru câteva perioade din viitor, pe baza observaţiilor din trecut. Istoria acestui proces este conţinută într-o mulţime a informaţiilor iar predicţia va fi bazată pe această mulţime de informaţii. În practică, această mulţime este o mulţime finită $\{y_n, y_{n-1}, ..., y_{n-p}\}$, reprezentând trecutul recent. Totuşi, în dezvoltarea teoriei predicţiei, este util a se considera o mulţime a informaţiilor infinită $I_n = \{y_n, y_{n-1}, ..., y_{n-p}, ...\}$, reprezentând întregul trecut.

Vom nota predicţia lui y_{n+h}, care este făcută la momentul n, prin $y_n(h)$ sau $\widehat{y}_{n+h}$.

Vom spune că **n este originea**, **h este orizontul de timp** iar $y_n(h)$ este **valoarea previzionată pentru** y_{n+h} **cu h perioade înainte**. În general, predictorul $y_n(h)$ (predictorul lui y_{n+h}, construit la timpul t), este o funcţie de variabilele din mulţimea I_t.

Definim **eroarea de previziune** (de prognoză) a lui y_{n+h}, diferenţa dintre valoarea reală şi cea previzionată: $e_n(h) = y_{n+h} - y_n(h) = y_{n+h} - \widehat{y}_{n+h}$. Ne aşteptăm ca valoarea medie a acestei erori să fie: $E((y_{n+h} - \widehat{y}_{n+h}) \mid I_n) = E(y_{n+h}) - \widehat{y}_{n+h} = \widehat{y}_{n+h} - \widehat{y}_{n+h} = 0$.

Varianţa erorii de prognoză $Var(e_n(h))$ este $E((y_{n+h} - \widehat{y}_{n+h})^2 \mid I_n)$. Acesta este necesară pentru previziunile pe interval de încredere, care sunt mai utile decât o estimare punctuală.

Criteriul folosit în mod obişnuit pentru a aprecia performanţa unui estimator sau predictor al unei variabile aleatoare este minimizarea pătratului erorii medii condiționate, de prognoză, adică: $\min E\left((y_{n+h} - y_n(h))^2 \mid I_n\right)$.

Vom folosi şi una din notaţiile: $y_n(h) = E(y_{n+h} \mid I_n) = E_t(y_{n+h})$.

$y_n(h)$ (sau $\hat{y}_{n+h}$) va fi calculat din ecuaţia modelului, scrisă pentru y_{n+h}, în felul următor:

– Înlocuim toţi parametrii necunoscuţi prin valorile lor estimate.

– Înlocuim variabilele aleatoare $y_1, y_2, ..., y_n$ prin valorile lor observate.

– Înlocuim variabilele aleatoare $y_{n+1}, y_{n+2}, ..., y_{n+h-1}$ prin valorile lor predictate: $\hat{y}_{n+1}, \hat{y}_{n+2}, ..., \hat{y}_{n+h-1}$

– Înlocuim variabilele eroare $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ prin reziduuri

– Înlocuim variabilele eroare $\varepsilon_{n+1}, \varepsilon_{n+2}, ..., \varepsilon_{n+h-1}$ prin 0 , adică media lor aşteptată.

Predicţia cu ajutorul modelului MA(q)

Un proces MA(q) are o memorie doar de lungime q şi aceasta limitează orizontul de prognoză.

De exemplu, presupunem că am estimat un model MA(2):

$$y_n = \mu + \varepsilon_n + \theta_1 \varepsilon_{n-1} + \theta_2 \varepsilon_{n-2}$$

Presupunem că ne bazăm pe *n* observaţii. Deoarece se presupune că parametrii sunt constanţi în timp, dacă relaţia anterioară are loc la momentul *n*, ea va avea loc şi la momentele $n+1, n+2, \ldots$. Vom putea scrie relaţiile:

$$y_{n+1} = \mu + \varepsilon_{n+1} + \theta_1 \varepsilon_n + \theta_2 \varepsilon_{n-1}$$

$$y_{n+2} = \mu + \varepsilon_{n+2} + \theta_1 \varepsilon_{n+1} + \theta_2 \varepsilon_n$$

$$y_{n+3} = \mu + \varepsilon_{n+3} + \theta_1 \varepsilon_{n+2} + \theta_2 \varepsilon_{n+1}$$

Ţinem cont că toate informaţiile până la momentul *n*, inclusiv cele de la la momentul *n*, sunt cunoscute şi disponibile. Realizarea de prognoze înseamnă a considera mediile condiţionate de informaţiile disponibile. Perturbaţiile $\varepsilon_{n+1}, \varepsilon_{n+2}, \ldots$ nu sunt cunoscute la momentul *n*. Cea mai bună prognoză pentru media condiţionată a lui ε_{n+1} este zero, adică $E_n(\varepsilon_{n+1}) = E(\varepsilon_{n+1} \mid I_n) = 0$, datorită ipotezelor procesului de zgomot alb. Vom obţine:

$$h = 1 \Rightarrow \hat{y}_{n+1} = y_n(1) = E(y_{n+1} \mid I_n) = E(\mu + \varepsilon_{n+1} + \theta_1 \varepsilon_n + \theta_2 \varepsilon_{n-1})$$

$$= \mu + E(\varepsilon_{n+1}) + \theta_1 E(\varepsilon_n) + \theta_2 E(\varepsilon_{n-1}) = \mu + \theta_1 \varepsilon_n + \theta_2 \varepsilon_{n-1}$$

$$e_n(1) = y_{n+1} - \hat{y}_{n+1} = \varepsilon_{n+1}$$

$$Var((e_n(1)) = Var(\varepsilon_{n+1}) = \sigma^2$$

Un interval de încredere 95% este de forma: $\hat{y}_{n+1} \pm 2\sigma$

$$h = 2 \Rightarrow \hat{y}_{n+2} = y_n(2) = E(y_{n+2} \mid I_n) = E(\mu + \varepsilon_{n+2} + \theta_1 \varepsilon_{n+1} + \theta_2 \varepsilon_n)$$

$$= \mu + E(\varepsilon_{n+2}) + \theta_1 E(\varepsilon_{n+1}) + \theta_2 E(\varepsilon_n) = \mu + \theta_2 \varepsilon_n$$

$$e_n(2) = y_{n+2} - \hat{y}_{n+2} = \varepsilon_{n+2} + \theta_1 \varepsilon_{n+1}$$

$$Var(e_n(2)) = Var(\varepsilon_{n+2} + \theta_1 \varepsilon_{n+1}) = (1 + \theta_1^2)\sigma^2$$

Un interval de încredere 95% este de forma: $\hat{y}_{n+1} \pm 2\sqrt{(1 + \theta_1^2)\sigma^2}$

$$h = 3 \Rightarrow \hat{y}_{n+3} = y_n(3) = E(y_{n+3} \mid I_n) = E(\mu + \varepsilon_{n+3} + \theta_1 \varepsilon_{n+2} + \theta_2 \varepsilon_{n+1})$$

$$= \mu + E(\varepsilon_{n+3}) + \theta_1 E(\varepsilon_{n+2}) + \theta_2 E(\varepsilon_{n+1}) = \mu$$

$$e_n(3) = y_{n+3} - \hat{y}_{n+3} = \varepsilon_{n+3} + \theta_1 \varepsilon_{n+2} + \theta_2 \varepsilon_{n+1}$$

$$Var(e_n(3)) = Var(\varepsilon_{n+3} + \theta_1 \varepsilon_{n+2} + \theta_2 \varepsilon_{n+1}) = (1 + \theta_1^2 + \theta_2^2)\sigma^2$$

Deoarece procesul MA(2) are o memorie de numai 2 perioade, toate predicţiile pentru 3 sau mai multe perioade în viitor, se reduc la parametrul de interceptare din model. Dacă în model nu există constantă, aceste prognoze vor fi 0.

Am folosit faptul că: $E_n(\varepsilon_{n+j}) = E(\varepsilon_{n+j} \mid I_n) = \begin{cases} \varepsilon_{n+j}, & j \le 0 \\ 0, & j > 0 \end{cases}$

Pentru $j \le 0$ avem valori cunoscute iar pentru $j > 0$ avem valori viitoare care au media 0.

Predicția cu ajutorul modelului AR(p)

Un proces AR are memorie infinită, spre deosebire de un proces MA(q) care are memorie de numai q perioade. Considerăm că a fost estimat un model AR(2): $y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$

Vom apela, din nou, la presupunerea de stabilitate a parametrilor şi vom putea scrie relațiile:

$y_n = c + \phi_1 y_{n-1} + \phi_2 y_{n-2} + \varepsilon_n$

$y_{n+1} = c + \phi_1 y_n + \phi_2 y_{n-1} + \varepsilon_{n+1}$

$y_{n+2} = c + \phi_1 y_{n+1} + \phi_2 y_n + \varepsilon_{n+2}$

$y_{n+3} = c + \phi_1 y_{n+2} + \phi_2 y_{n+1} + \varepsilon_{n+3}$.

Obținerea prognozei pentru o perioadă în viitor este simplă deoarece toate informațiile despre y sunt cunoscute la timpul *n*. Aplicăm operatorul de medie şi considerăm $E_n(\varepsilon_{n+1}) = E(\varepsilon_{n+1} \mid I_n) = 0$.

$h = 1 \Rightarrow \hat{y}_{n+1} = y_n(1) = E(y_{n+1} \mid I_n) = E(c + \phi_1 y_n + \phi_2 y_{n-1} + \varepsilon_{n+1}) = c + \phi_1 y_n + \phi_2 y_{n-1}$

$$e_n(1) = y_{n+1} - \hat{y}_{n+1} = \varepsilon_{n+1} \Rightarrow Var((e_n(1)) = Var(\varepsilon_{n+1}) = \sigma^2$$

Un interval de încredere 95% este de forma: $\hat{y}_{n+1} \pm 2\sigma$

$h = 2 \Rightarrow \hat{y}_{n+2} = y_n(2) = E(y_{n+2} \mid I_n) = E(c + \phi_1 y_{n+1} + \phi_2 y_n + \varepsilon_{n+2}) = c + \phi_1 \hat{y}_{n+1} + \phi_2 y_n$

$$e_n(2) = y_{n+2} - \hat{y}_{n+2} = c + \phi_1 (y_{n+1} - \hat{y}_{n+1}) + \varepsilon_{n+2} = c + \phi_1 \varepsilon_{n+1} + \varepsilon_{n+2}$$

$$Var(e_n(2)) = Var(c + \phi_1 \varepsilon_{n+1} + \varepsilon_{n+2}) = (1 + \phi_1^2)\sigma^2$$

Un interval de încredere 95% este de forma: $\hat{y}_{n+1} \pm 2\sqrt{(1 + \phi_1^2)\sigma^2}$

$h = 3 \Rightarrow \hat{y}_{n+3} = y_n(3) = E(y_{n+3} \mid I_n) = E(c + \phi_1 y_{n+2} + \phi_2 y_{n+1} + \varepsilon_{n+3}) = c + \phi_1 \hat{y}_{n+2} + \phi_2 \hat{y}_{n+1}$

$$e_n(3) = y_{n+3} - \hat{y}_{n+3} = c + \phi_1 y_{n+2} + \phi_2 y_{n+1} + \varepsilon_{n+3} - (c + \phi_1 \hat{y}_{n+2} + \phi_2 \hat{y}_{n+1}) =$$
$$= \phi_1 (y_{n+2} - \hat{y}_{n+2}) + \phi_2 (y_{n+1} - \hat{y}_{n+1}) + \varepsilon_{n+3}$$

Predicția cu ajutorul unui model ARMA(2,2)

Considerăm un model ARMA(2,2): $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$

Dorim o predicție la momentul n, peste 2 perioade în viitor, deci pentru h=2.

$y_n = \phi_1 y_{n-1} + \phi_2 y_{n-2} + \varepsilon_n + \theta_1 \varepsilon_{n-1} + \theta_2 \varepsilon_{n-2}$

La momentul $n+1$ avem ecuația: $y_{n+1} = \phi_1 y_n + \phi_2 y_{n-1} + \varepsilon_{n+1} + \theta_1 \varepsilon_n + \theta_2 \varepsilon_{n-1}$

La momentul $n+2$ avem ecuația: $y_{n+2} = \phi_1 y_{n+1} + \phi_2 y_n + \varepsilon_{n+2} + \theta_1 \varepsilon_{n+1} + \theta_2 \varepsilon_n$

$\hat{y}_{n+2} = y_n(2) = E(y_{n+2} \mid I_n) = \phi_1 \hat{y}_{n+1} + \phi_2 y_n + \theta_2 \varepsilon_n$

Predicția cu ajutorul modelului ARMA(p,q)

Prognozele pot fi generate prin aşa numita funcție de prognoză:

$f_{n,h} = \sum_{i=1}^{p} \phi_i f_{n,h-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{n+h-j}$,unde $f_{n,h} = y_{n+h}, h \le 0$, $\varepsilon_{n+h} = \varepsilon_{n+h}, h \le 0$ şi $\varepsilon_{n+h} = 0, h > 0$.

Predicția folosind modele ARIMA(p,d,q)

Presupunem că seria $y_1, y_2, \ldots, y_n$ urmează un model general ARIMA(p,d,q), model ce poate fi rescris în funcție de valoarea curentă şi de valorile anterioare ale lui ε_t:

$$y_t = \frac{\Theta(L)}{\Phi(L)\Delta^d} \varepsilon_t = \Psi_\infty(L)\varepsilon_t = (1 + \psi_1 L + \psi_2^2 L^2 + \cdots)\varepsilon_t$$

Valoarea viitoare y_{n+h} este generată prin model astfel:

$y_{n+h} = \varepsilon_{n+h} + \psi_1 \varepsilon_{n+h-1} + \psi_2 \varepsilon_{n+h-2} + \cdots$

Prognoza optimală a lui y_{n+h} este media condiționată $y_n(h) = E(y_{n+h} \mid I_n) = f_{n,h}$. Termenul de prognoză optimală este folosit în sensul că minimizează eroarea pătratică medie (MSE).

Dacă procesul este normal, un interval de încredere $(1-\alpha)$ va fi $[y_n(h) \pm z_{\alpha/2}\sqrt{Var(e_n(h))}]$. Pentru $h = 1$ avem $e_n(1) = y_{n+1} - y_n(1) = \varepsilon_{n+1}$, deci σ_ε^2 este varianța erorii de prognoză pentru o perioadă în viitor.