## 3.5. Системний підхід до побудови регресійної моделі по часових рядах

Відомі методики побудови моделей типу авторегресії з ковзним середнім (АРКС), АРКС з ендогенними змінними (АРКСЕ) або АРКС з інтегрованим ковзним середнім (АРІКС) . Однак у представлених методиках нечітко представлене поняття структури моделі, а також недостатньо уваги приділяється визначенню нелінійностей моделі. Пропонований нижче системний підхід може бути використаний при побудові лінійних моделей, а також моделей з нелінійностями щодо змінних (псевдолінійні моделі). Хоча моделі, нелінійні щодо параметрів, тут розглядатись не будуть, окремі елементи пропонованої методики можуть бути застосовані також при побудові моделей і такого класу.

Відповідно до пропонованого підходу побудова моделі за часовими рядами складається з п’яти наступних етапів [3]:

- Виконати аналіз процесу (процесів), для якого будується модель на підставі вимірів вхідних і вихідних змінних, представлених відповідними часовими рядами.
- Виконати аналіз наявних часових рядів на можливу присутність нелінійностей за допомогою ряду критеріїв.
- Вибрати структури моделей-кандидатів, для чого необхідно виконати наступне: обчислити і виконати аналіз кореляційної матриці для часових рядів залежної і незалежної змінних з метою визначення екзогенних змінних, котрі необхідно включити в модель; обчислити автокореляційну і приватну автокореляційну функцію для залежної змінної з метою вибору порядку авторегресійної частини моделі.
- Вибрати метод (методи) для оцінювання коефіцієнтів (параметрів) моделей-кандидатів і оцінити їх параметри.
- Вибрати кращу (адекватну) модель з отриманого на четвертому етапі множини кандидатів, використовуючи для цієї мети набір статистичних параметрів.

**Структура моделі**. Перш ніж перейти до розгляду конкретних етапів побудови моделі, розглянемо поняття структури математичної моделі, що буде використовуватися надалі. Поняття структури моделі містить у собі наступне:

1. *Порядок моделі*, тобто порядок диференційного, різницевого або іншого рівняння, яке використовується для опису динаміки про-

цесу або об'єкта. Наприклад, стохастичне різницеве авторегресійне (АР) рівняння другого порядку має вигляд:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \varepsilon(k) \qquad (3.36)$$

Тобто порядок цього різницевого рівняння визначається числом затриманих у часі значень змінної, які використовуються у правій частині рівняння. Стохастичним воно називається тому, що в правій частині присутня випадкова змінна $\varepsilon(k)$, призначення якої розглянемо нижче. Слід зазначити, що введення випадкової складової обов'язково вимагає опису її основних (передбачуваних або відомих точно) статистичних характеристик, таких як математичне очікування, дисперсія, автокореляційна функція і корельованість з ендогенної змінної.

2. *Розмірність моделі*. Вона визначається числом рівнянь, які використовуються для опису об'єкта або процесу. Процес, що описують одним рівнянням, називають одномірним або скалярним. Процес, що описують двома і більше рівняннями, називають багатомірним. Зручним є представлення в просторі станів. При цьому розмірність моделі відповідає розмірності вектора стану процесу (об'єкта).

3. *Наявність нелінійностей і їх характер*. Визначити наявність нелінійностей – не завжди проста задача. Так, для механічних і деяких інших систем наявність нелінійностей можна визначити шляхом попереднього вивчення законів, закономірностей і особливостей їх функціонування. Наприклад, відомо, що для механічних систем характерними є нелінійності типу "люфт", "тертя", білінійності, а для електричних – гістерезис.

При побудові регресійних моделей найчастіше зустрічаються нелінійності щодо змінних і нелінійності щодо параметрів. Прикладом нелінійності щодо змінних може бути розповсюджена поліноміальна стохастична регресія вигляду

$$y(k) = a_0 + a_1 x(k) + a_2 x^2(k) + a_3 x^3(k) + \varepsilon(k) \qquad (3.37)$$

Коефіцієнти цього рівняння можна оцінювати звичайним методом найменших квадратів (МНК) при належній побудові матриці вимірів. Ще одним прикладом може бути логістичне рівняння

$$y(k) = a\, y(k-1) - a\, y^2(k-1) + \varepsilon(k) \qquad (3.38)$$

яке описує нелінійні процеси при $0 < a \le 4$ і $y(0) \in (0,1)$. У граничному випадку (при $a = 4$) це рівняння описує хаотичний процес.

Нелінійність по параметрах обумовлена наявністю в моделі добутків коефіцієнтів, наприклад, у вигляді

$$y(k) = a_0 + a_1 a_2 x(k) + a_2 \exp(-bx(k)) + \varepsilon(k)$$

Коефіцієнти (параметри) такої моделі неможливо оцінити за допомогою звичайного МНК, тому для рішення цієї задачі використовують нелінійний МНК, метод максимальної правдоподібності або інші методи нелінійного оцінювання.

4. *Час запізнювання* реакції на виході об'єкта стосовно вхідного сигналу. Запізнювання по входу, якщо воно відомо, досить легко враховується як у безперервних, так і в дискретних моделях. Для дискретної моделі у вигляді різницевого рівняння

$$y(k) = a_0 + a_1 y(k-1) + a_2 x(k-d) + \varepsilon(k) \qquad (3.39)$$

час запізнювання $d$ являє собою ціле число, рівне кількості періодів дискретизації вимірів, на які вихідний сигнал запізнюється щодо вхідного, тобто $d = \text{int}[\tau / T_s]$, де $\tau$ величина запізнювання в безперервному часі; $T_s$ – період дискретизації вимірів. Тривалість періоду дискретизації вимірів залежить від динаміки конкретного процесу і може змінюватися в межах від декількох мікросекунд для фізико-технічних систем до одного року в макроекономіці.

5. *Тип збурювань*, що діють на процес, і спосіб їх обліку. Під збуреннями розуміють вхідні впливи процесу, що роблять, як правило, негативний вплив на його протікання, але не використовуються як керуючі. Збурювання поділяють на детерміновані і стохастичні, а враховуються вони в адитивній або мультиплікативній формі. Вище ми навели різницеві рівняння, у які збурювання $\varepsilon(k)$ входять в адитивній формі. Приклад мультиплікативної форми:

$$h(k) = \upsilon(k)[\alpha_0 + \alpha_1 h(k-1)] \qquad (3.40)$$

де $\upsilon(k)$ – мультиплікативне збурювання. Введення випадкової складової в модель обумовлено наступними основними причинами: присутність неконтрольованих зовнішніх збурювань, введення в модель зайвих пояснюючих змінних або, навпаки, відсутність у моделі необхідних пояснюючих змінних, вплив методичних і обчислювальних погрішностей.

Вибір структури моделі, адекватної процесові, – задача досить не проста і вирішується, як правило, ітеративно. Спочатку структуру

моделі оцінюють приблизно на підставі аналізу відомої інформації про процес, дослідження закономірностей його протікання, аналізу кореляційних функцій, візуального аналізу даних. При цьому доцільно вибирати трохи найбільш ймовірні структури (кандидатів). Потім визначають оцінки параметрів моделей-кандидатів і вибирають кращу з них, використовуючи відповідні статистичні характеристики моделей.

Якщо жодна з моделей-кандидатів не може вважатися адекватною, то необхідно досліджувати на інформативність експериментальні дані, що можуть бути недостатньо інформативними для оцінювання моделі. У такому випадку може знадобитися повторний або додатковий збір експериментальних даних.

**Аналіз процесу**. На цьому етапі необхідно скористатися всією наявною інформацією про процес з метою визначення числа його входів і виходів; логічних взаємозв'язків між змінними; можливої присутності нелінійностей і їх характеру; визначення типу збурювань, що діють на процес; визначення присутності запізнювання на якісному і, можливо, кількісному рівнях; приблизного визначення порядку процесу. У випадку дослідження економічних процесів необхідно встановити, чи мається вплив сезонних ефектів, чи є присутнім тренд (на якісному рівні); можливо, що виникне необхідність висунути гіпотезу про існування випадкового тренда; чи є ділянки часових рядів з істотно різними рівнями коливань (присутність гетероскедастичності); оцінити необхідність використання гіпотези відносно коінтегрованості змінних. У результаті аналізу процесу необхідно в загальному вигляді постулювати структуру математичної моделі, що буде використовуватися надалі для опису його поведінки. Наприклад, якщо висувається гіпотеза про існування гетероскедастичності, то необхідно вибрати можливий клас моделей для її опису. Те ж саме стосується присутності коінтегрованості змінних або випадкового тренда.

**Визначення наявності нелінійностей**. Для рішення цієї задачі можна користуватися різними критеріями. Однак при цьому необхідно знати про їх можливості. Покажемо на простому прикладі, що застосування лінійних коваріаційних функцій не завжди приводить до позитивних результатів. Нехай при визначенні структури моделі не були враховані деякі пояснюючі змінні та у результаті корельовані залишки описуються наступним рівнянням:

$$\xi(k) = c\,u(k-1)\,e(k-1) + e(k) \tag{3.41}$$

де $e(k)$ – білий гауссовський шум; $E[e(k)]=0$, $E[u(k)]=0$, $E[e(k)u(k)]=0$, тобто змінні $e(k)$ і $u(k)$ некорельовані і мають нульове середнє; $c$ – масштабний коефіцієнт. Можна показати, що нормована автокореляційна функція залишків і нормована функція взаємної кореляції між вхідним сигналом $u(k)$ і залишками мають вигляд

$$\Phi_{\xi\xi}(\tau) = \delta(\tau), \qquad \Phi_{u\xi}(\tau) = 0, \quad \forall\tau \tag{3.42}$$

Однак з рівняння (1) випливає, що $\xi(k)$ – корельована послідовність, що буде вносити зсув в оцінки параметрів моделі. Таким чином, у загальному випадку лінійні кореляційні методи не дають можливості визначити факт присутності нелінійних ефектів і їх вплив на процес.

Для того, щоб оцінити тип зв'язку між входом і виходом (тобто зв'язок лінійний або нелінійний), можна скористатися спектральною функцією високого порядку вигляду:

$$X_{ij} = \frac{\left|S_{\omega}(\omega_i, \omega_j)\right|^2}{S_{\omega}(\omega_i) S_{\omega}(\omega_j) S_{\omega}(\omega_i / \omega_j)} \tag{3.43}$$

де $S_{\omega}(\omega_i,\omega_j)$ – біспектральна щільність потужності; $S_{\omega}(\omega_i)$ – спектральна щільність потужності часового ряду. При $S_w(\omega_i,\omega_j) = 0$, $\forall\omega_i,\omega_j$ процес буде лінійним і третім моментом вхідного сигналу $\mu_3 = 0$. Однак, якщо $X_{ij} = \text{const}$, то процес лінійний, але $\mu_3 \neq 0$.

Такий підхід до визначення присутності нелінійностей має два недоліки. По-перше, оцінювання спектральної щільності потужності вимагає застосування спеціальної попередньої обробки сигналів у вигляді застосування часових вікон, усереднення, цифрової фільтрації і т.п. По-друге, він не завжди може бути використаний при рішенні задач ідентифікації систем, оскільки він не дає можливості одержати оцінки параметрів моделі в явному вигляді. Крім того, при рішенні цих же задач не завжди є можливість одержати виміри вхідного сигналу або ж інформативний вхідний сигнал одержують штучно у вигляді спеціально генерованих послідовностей, що не завжди можна подавати на вхід об'єкта внаслідок особливостей його функціонування.

Що стосується економічних процесів, то в цьому випадку, як правило, не можна поставити експеримент із процесом. Тому використовують тільки ті статистичні дані, які можна реально зібрати в процесі дослідження. У загальному випадку при ідентифікації систем використовують три типи сигналів: вхідний, вихідний і збурений. При цьому вхідний керуючий сигнал вважають незалежним від збурювання. У результаті виявляється неможливим з'ясувати деякі типи зв'язків.

Можливе використання також дисперсійного методу визначення присутності нелінійностей, що заснований на застосуванні наступної функції:

$$\Psi_{zu}(t_1,t_2)=E_{u(t_2)}[E_{z(t_1)}[z(t_1)\,|\,u(t_2)]-E_{z(t_1)}[z(t_1)]]^2 \qquad (3.44)$$

яка обчислюється за допомогою досить складного інтегрального рівняння, якщо відомі відповідні щільності розподілу імовірностей сигналів, що не завжди можна визначити.

У зв'язку з вищесказаним для виявлення нелінійностей представляється доцільним використовувати більш прості кореляційні процедури. Нехай система представлена в аналітичній формі за допомогою рядів Вольтерра:

$$z(t)=\sum_{k=1}^{\infty}\int_{-\infty}^{+\infty}\ldots\int h_n(\tau_1,\tau_2,\ldots,\tau_n)\,\Pi_{i=1,\ldots,n}\,u(t-\tau_i)\,d\tau+e(t) \qquad (3.45)$$

Використовуючи операторне представлення, запишемо це рівняння у вигляді

$$z(t)=\sum_{n=1}^{\infty}H_n[u(t)]+e(t)=H[u(t)]+e(t)=$$

$$=\sum_{n=1}^{\infty}H_n(u^n(t))+e(t), \qquad (3.46)$$

де квадратні дужки вказують на те, що $H$ – це оператор для $u(t)$, а круглі дужки – на фактичну залежність.

Надалі будемо думати, що випадкові сигнали, що зустрічаються в процесі ідентифікації, є ергодичними, тобто середні значення по ансамблю можуть бути перетворені в середні за часом за допомогою деякої вибіркової функції. Розглянемо чутливість моделі Вольтерра другого порядку до вхідного сигналу $u(t) + b$. У даному випадку вихідний сигнал визначається як

$$z(t) = H_1[u(t) + b] + H_2[u(t) + b] + e(t) =$$

$$= H_1(u(t) + b) + H_2(u_2(t) + 2bu(t) + b^2) + e(t). \qquad (3.47)$$

Якщо відняти середнє з вихідної величини, то одержимо:

$$z'(t) = H_1(u(t)) + H_2(u^2(t) + 2bu(t) - \bar{u}^2(t)) + e'(t)$$

де штрихом позначений процес з нульовим середнім. Ця модель включає залежність від $\sigma_u^2 = u^2(t)$ і від $b$, тому вона буде давати правильний прогноз тільки в тому випадку, коли вхідний сигнал має таку ж характеристику. Таким чином, чутливість моделі до вхідного сигналу залежить від її типу, тобто від її структури.

Для того, щоб вихідний сигнал не залежав від дисперсії вхідного, віднімемо з останнього рівняння середнє при $u(t) = 0$, тобто величину

$$\bar{z}_b(t) = H_1[b] + H_2[b] + ... + e(t).$$

У результаті одержимо наступну залежність:

$$z'_b(t) = z(t) - \bar{z}_b(t) = H_1(u(t)) +$$

$$+ H_2(u^2(t) + 2bu(t)) + ... + e'(t) \qquad (3.48)$$

З (3.47) і (3.48) випливає, що $\bar{z}_b(t) = \bar{z}(t)$ тоді і тільки тоді, коли об'єкт лінійний, тобто останнє рівняння можна використовувати як простий тест на присутність нелінійності.

Задачу виявлення нелінійностей сформулюємо в такий спосіб: *потрібно встановити необхідність застосування нелінійної моделі для опису конкретної вибірки даних. Для рішення задачі будемо користуватися кореляційними функціями*.

Нехай вхідний сигнал $u(t)$ і шум $e(t)$ – незалежні процеси з нульовим середнім, і нехай усі моменти з непарними ступенями для цих сигналів дорівнюють нулеві, а для вхідного сигналу існують усі моменти з парними ступенями. Розглянемо кореляційну функцію $\Phi_{z'z'^2}(\tau)$, де $z'(t)$ – відгук системи на вхідний сигнал $u(t) + b$ після видалення з нього середнього значення. По визначенню кореляційна функція $\Phi_{z'z'^2}(\tau)$ визначається як

$$\Phi_{z'z'^2}(\tau) = E[z'(t+\tau)(z'(t))^2], \qquad (3.49)$$

де

$$z'(t+\tau) = \int h_1(\tau_1)(u(t-\tau_1+\tau)+b)\,d\tau_1 +$$

$$+\iint h_2(\tau_1,\tau_2)(u(t-\tau_1+\tau)+b(u(t-\tau_2+\tau)+b)\,d\tau_1\,d\tau_2+\ldots$$

$$+e(t+\tau). \qquad (3.50)$$

Після заміни змінних в останньому рівнянні одержимо

$$z'(t+\tau) = \int h_1(t-\tau_1+\tau)u(\tau_1)\,d\tau_1 +$$

$$+\iint h_2(t-\tau_1+\tau,t-\tau_2+\tau)(u(\tau_1)u(\tau_2)+bu(\tau_1)+bu(\tau_2))d\tau_1\,d\tau_2+\ldots$$

$$-\iint h_2(t-\tau_1+\tau,t-\tau_2+\tau)\overline{u}(\tau_1)\overline{u}(\tau_2)\,d\tau_1\,d\tau_2+\ldots+e'(t+\tau) \qquad (3.51)$$

З урахувунням (3.46) останнє рівняння запишемо у вигляді

$$z'(t+\tau)=H_1^{\tau}(u(t))+H_2^{\tau}(u^2(t))+2bH_2^{\tau}(u(t))-$$

$$-H_2^{\tau}(\overline{u}^2(t))+\ldots+e'(t+\tau). \qquad (3.52)$$

Тепер функція (3.49) приймає вигляд:

$$\Phi_{z'z'^2}(\tau) = E[z'(t+\tau)(z'(t))^2] =$$

$$E\left\{\begin{matrix}\left[H_1(u)+H_2(u^2+2bu-\overline{u}^2)+H_3(u^3+3bu^2+3b^2u-3b\overline{u}^2+\ldots+e'(t)\right]^2\times \\ \left[H_1^{\tau}(u)+H_2^{\tau}(u^2+2bu-\overline{u}^2)+H_3^{\tau}(u^3+3bu^2+3b^2u-3\overline{b}\overline{u}^2+\ldots+e'(t+\tau)\right]\end{matrix}\right\}=$$

$$=E\{[(H_1H_1)(u^2)+2(H_1H_2)(u^3+2bu^2-\overline{u}\overline{u}^2)+2(H_1H_3)(u^4+3bu^3+3b^2u^2-$$

$$-3b\overline{u}^2u)+\ldots+e'^2(t)]\times$$

$$\times[H_1^{\tau}(u)+H_2^{\tau}(u^2+2bu-\overline{u}^2)+H_3^{\tau}(u^3+3bu^2+$$

$$+3b^2u-3\overline{b}\overline{u}^2)+\ldots+e'(t+\tau)]\} \qquad (3.53)$$

Виконаємо аналіз кореляційної функції $\Phi_{z'z'^2}(\tau)$. Розглянемо окремо кожен член рівняння (3.53) з врахуванням того, що всі непарні моменти вхідного сигналу дорівнюють нулеві, а парні – присутні. У результаті одержуємо:

(а)

$$E[(H_1H_1)(u^2)H_1^{\tau}(u)] = E[(H_1H_1H_1^{\tau})(u^3)] = 0 \qquad (3.54)$$

(б)

$$E[(H_1H_1)(u^2)H_2^{\tau}(u^2 + 2bu - \bar{u}^2)] = E[(H_1H_1H_2^{\tau})$$

$$(u^4 + 2bu^3 - \bar{u}^2u^2)] \neq 0 \qquad (3.55)$$

(в)

$$E[(H_1H_1(u^2)H_3^{\tau}(u^3 + 3bu^2 + 3b^2u - 3b\bar{u}^2)] =$$

$$= E[(H_1H_1H_3^{\tau})(u^5 + 3bu^4 + 3b^2u^3 - 3b\bar{u}^2u^2)] \neq 0 \qquad (3.56)$$

(г)

$$E[(2H_1H_2)(u^3 + 2bu^2 - u\bar{u}^2)H_1^{\tau}(u) =$$

$$= E[(2H_1H_2H_1^{\tau})(u^4 + 2bu^3 - \bar{u}^2u^2)] \neq 0 \qquad (3.57)$$

За аналогією можна показати, що всі інші члени (за винятком тих, що містять сигнал помилки $e(t)$ також не дорівнюють нулеві і впливають на значення кореляційної функції. Нульові функції мають вигляд

$$E[(H_1H_1)(u^2)e'(t+\tau)] = 0$$

$$E[(2H_1H_2)(u^3 + 2bu^2 - \bar{u}^2u)e'(t+\tau) = 0,$$

$$\cdots$$

$$E[(e'^2)e'(t+\tau) = 0$$

З наведеного аналізу випливає, що

$$\Phi_{z'z'^2}(\tau) = 0, \qquad \forall\tau \qquad (3.58)$$

тоді і тільки тоді, коли об'єкт лінійний, тобто $H_2, H_3,\ldots, H_4 = 0$. Таким чином, об'єкт буде містити нелінійності, коли $\Phi_{z'z'^2}(\tau) \neq 0$. Гіпотеза щодо рівності нулеві третього моменту вхідного сигналу виконується при порушенні об'єкта рівномірно розподіленим гауссовським шумом і іншими випадковими процесами. Вона перевіряється за допомогою наступної коваріаційної функції:

$$E[u(t)u(t+\tau_1)u(t+\tau_1)], \quad \forall \tau_1, \tau_2$$

Присутність у вхідному сигналі постійної $b$ сприяє виявленню нелінійностей системи, що впливають на величину $\Phi_{z'z'^2}(\tau)$. Якщо покласти $b = 0$, то третій член розкладання (3.54)-(3.57) буде дорівнювати нулеві і за допомогою функції $\Phi_{z'z'^2}(\tau)$ буде неможливо визначити нелінійності непарного порядку. При наявності вимірів величини $z_b'(t)$ результат, подібний (3.58), можна одержати також для функції $\Phi_{z_b z_b^2}(\tau)$.

Крім розглянутих підходів до визначення наявності нелінійностей при побудові регресійних моделей, можна скористатися більш простими тестами. Наприклад, статистикою

$$\widehat{F} = \frac{\dfrac{1}{k-2}\sum_{i=1}^{k} n_i (\bar{y}_i - \hat{y}_i)^2}{\dfrac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}$$

де $k$ – число груп даних; $n_i$ – число вимірів у групі; $n$ – загальне число вимірів. Фактично дана статистика являє собою наступне відношення:

$$\widehat{F} = \frac{\textit{Відхилення середніх значень від прямої регресії}}{\textit{Відхилення значень } y(k) \textit{ від групових середніх}}.$$

Якщо статистика $\widehat{F}$ з $\nu_1 = k-2, \nu_2 = n-k$ ступенями волі досягає або перевершує рівень значимості, то гіпотезу про лінійність потрібно відкинути.

**Вибір структури моделей-кандидатів**. *Коефіцієнт кореляції*, а в загальному випадку – *кореляційна функція*, дозволяє встановити наявність зв'язку між ендогенними (залежними) і екзогенними (незалежними) змінними. Кореляція може бути лінійною або нелінійною в залежності від типу залежності, що фактично існує між змінними. У більшості практичних випадків розглядають лінійну кореляцію

(взаємозв'язок), однак більш глибокий аналіз вимагає залучення для дослідження процесів нелінійних залежностей. Складну нелінійну залежність можна спростити, але знати про її існування необхідно для того, щоб побудувати адекватну модель процесу.

Кореляційна матриця дозволяє встановити факт наявності зв'язку між зазначеними змінними. Розглянемо кореляційну матрицю розмірності 3 × 3, що будується для трьох змінних $x$, $y$, $z$:

$$R = \begin{bmatrix} r_{yy} & r_{xy} & r_{zy} \\ r_{yx} & r_{xx} & r_{zx} \\ r_{yz} & r_{xz} & r_{zz} \end{bmatrix} \tag{3.59}$$

де $r_{yx} = r_{xy}$, $r_{yz} = r_{zy}$, $r_{xz} = r_{zy}$.

Нехай $y$ – залежна змінна, а $x$, $z$ – технологічні параметри, що приблизно впливають на $y$. Тобто ми визначаємо наявність залежності вигляду:

$$y = f(x, z)$$

яка може бути представлена у формі регресії змінної $y$ на незалежні змінні $x$, $z$:

$$y(k) = a_0 + a_1 x(k) + a_2 z(k) + \varepsilon(k) \tag{3.60}$$

де $k$ – дискретний час (наприклад, у секундах, хвилинах, годинах, днях, тижнях, місяцях і т.д.); $\varepsilon(k)$ – випадкова змінна, причини введення якої в модель були розглянуті вище.

Найчастіше вважають, що сукупний вплив усіх зазначених факторів можна з деяким допущенням описати випадковою змінною $\varepsilon(k)$. Оскільки вона не вимірюється, то оцінити її значення (помилку моделі або залишок) можна тільки після оцінювання коефіцієнтів моделі, тобто

$$\varepsilon(k) \approx e(k) = \hat{y}(k) - y(k)$$

де $\hat{y}(k)$ – оцінка змінної $y(k)$, отримана по моделі; $y(k)$ – вимір.

Для обчислення елементів матриці $R$ необхідно мати синхронні за часом вибірки значення усіх трьох змінних $y$, $x$, $z$. Формула для розрахунку коефіцієнтів кореляції має вигляд

$$r_{yx} = \frac{1}{N} \frac{\sum_{k=1}^{N} \{[x(k) - \bar{x}][y(k) - \bar{y}]\}}{\sigma_x \sigma_y}$$

де $\bar{x}, \bar{y}-$ середні вибіркові значення змінних $x$, $y$; $\sigma_x$, $\sigma_y$ – стандартні відхилення цих змінних, тобто

$$\sigma_y = \sqrt{\sigma_y^2} = \left[ \frac{1}{N-1} \sum_{k=1}^{N} [y(k) - \bar{y}]^2 \right]^{1/2}$$

де $N$ – число вимірів змінної $y$.

Коефіцієнти кореляції показують ступінь взаємозв'язку між змінними. Очевидно, що перш ніж формально обчислювати коефіцієнти кореляції, необхідно виконати аналіз процесу і визначити присутність (або відсутність) логічного зв'язку між змінними. Це дозволяє ввести в розгляд тільки ті змінні, котрі дійсно впливають на залежну. Очевидно, що для правильного вибору змінних необхідно досить глибоко знати модельований процес (для рішення цієї задачі введено перший етап).

На підставі значень коефіцієнтів кореляції приймається рішення про включення їх у рівняння регресії:

$$y(k) = a_0 + b_1 x(k) + b_2 z(k) + \varepsilon(k)$$

яке може бути представлене в загальному вигляді як множинна регресія

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + a_3 x_3(k) + \ldots$$
$$\ldots + a_{p-1} x_{p-1}(k) + \varepsilon(k) \qquad (3.61)$$

Відомо, що між коефіцієнтами регресії $b_1$, $b_2$ і коефіцієнтами кореляції $r_{yx}$, $r_{yz}$ існує однозначний взаємозв'язок.

Рівняння (3.61) являє собою *множинну лінійну регресію* $p$-го порядку, хоча найчастіше приходиться застосовувати більш складні нелінійні моделі. Характерним представником нелінійної за змінними регресії є поліноміальна регресія порядку.

Для визначення необхідності включення в рівняння регресії авторегресійної складової необхідно обчислити і досліджувати вибіркову *автокореляційну і приватну автокореляційну функцію* змінної $y(k)$.

Рівняння з авторегресійної складової має вигляд

$$y(k) = a_0 + a_1 y(k-1) + \\ + a_2 y(k-2) + b_1 x(k) + b_2 z(k) \qquad (3.62)$$

тобто у рівняння регресії додана авторегресійна (АР) складова другого порядку. Порядок авторегресії визначається за допомогою автокореляційної функції. Число коефіцієнтів автокореляційної функції, що відмінні від нуля в статистичному змісті, і буде складати порядок авторегресії.

Коефіцієнти автокореляційної функції обчислюють за формулою:

$$r_y(s) = r_{y(k)y(k-s)} = \frac{1}{N} \frac{\sum_{k=s+1}^{N} \{[y(k) - \bar{y}][y(k-s) - \bar{y}]\}}{\sigma_y^2},$$

$$s = 1,2,3,... \qquad (3.63)$$

де $\sigma_y^2$ – вибіркова дисперсія змінної $y(k)$. Число коефіцієнтів АКФ, відмінних від нуля в статистичному змісті, вказує на порядок авторегресійної частини моделі.

Уточнити порядок авторегресійної складової дозволяє приватна автокореляційна функція (ПАКФ), що обчислюється відповідно до виразів:

$$\Phi_{11} = r(1) \qquad \Phi_{22} = \frac{r_2 - r_1^2}{1 - r_1^2}$$

$$\Phi_{ss} = \frac{r_s - \sum_{j=1}^{s-1} \Phi_{s-1,j} r_{s-j}}{1 - \sum_{j=1}^{s-1} \Phi_{s-1,j} r_j} \qquad (3.64)$$

ПАКФ чіткіше відображає порядок АР-моделі завдяки відсутності впливу проміжних коефіцієнтів кореляції на обрані значення змінної, тобто коефіцієнт $\Phi_{11}$ характеризує ступінь взаємозв'язку між стоячими поруч (за часом) значеннями змінної, а $\Phi_{22}$ характеризує взаємоз-

в'язок між значеннями змінної, віддаленими на відстані двох періодів дискретизації.

Значення коефіцієнтів вибіркової (тобто обчисленої по вибірці експериментальних даних) приватної автокореляційної функції можна приблизно визначити за експериментальними даними у такий спосіб. Коефіцієнт $a_{11}$ моделі

$$y(k) = a_{11} y(k-1)$$

можна поставити у відповідність коефіцієнту ПАКФ $a_{11} \approx \Phi_{11}$, а коефіцієнт $a_{22}$ моделі

$$y(k) = a_{22} y(k-2)$$

приблизно дорівнює коефіцієнтові $\Phi_{22}$. Коефіцієнти $a_{11}$, $a_{22}$ оцінюють, наприклад, методом найменших квадратів.

Коли ми говоримо, що значення коефіцієнтів автокореляційної функції повинні бути відмінними від нуля в статистичному змісті, це означає, що існує деякий вираз, що дозволяє встановити або спростувати цей факт. Одним із загальноприйнятих підходів до визначення того факту, що коефіцієнти АКФ істотно відмінні від нуля в статистичному змісті, є обчислення статистичного параметра (або просто статистики) Льюнга-Бокса $Q(r_k)$, що обчислюється за формулою

$$Q(r_k) = N(N+2) \sum_{k=1}^{s} r_k^2 \,/(N-k)$$

де $N$ – довжина вибірки даних змінної, для якої знайдені значення автокореляційної функції $r_k$; $s$ – число коефіцієнтів АКФ, що досліджуються, на істотну відмінність від нуля.

Третій етап закінчується вибором структур декількох моделей-кандидатів, коефіцієнти яких будуть оцінюватися на наступному етапі.

**Оцінювання коефіцієнтів моделей-кандидатів.** На цьому етапі обчислюють оцінки коефіцієнтів моделей-кандидатів, що розрізняються своєю структурою. Наприклад, можна вибрати авторегресійну частину (модель) першого, другого і третього порядку. Можна розглянути моделі, що окремо включають пояснюючі змінні, а також моделі, що містять усі пояснюючі змінні разом. Найбільш розповсюдженими методами оцінювання параметрів моделі є наступні: метод найменших квадратів (МНК) і його модифікації; метод максимальної

правдоподібності (ММП); метод допоміжної змінної (МДЗ); нелінійний метод найменших квадратів (НМНК) і їхні рекурсивні версії.

Для одержання незміщених оцінок вектора параметрів θ регресійної моделі за допомогою методу найменших квадратів необхідно виконати наступні умови:

а) ε(*k*) некорельована послідовність випадкових чисел з нульовим середнім, тобто

$$E[\varepsilon(k)] = 0 \qquad \operatorname{cov}[\varepsilon(k)] = E[\varepsilon(k)\varepsilon(j)] = \begin{cases} \sigma_\varepsilon^2, & k = j; \\ 0, & k \neq j. \end{cases}$$

б) послідовності ε(*k*) і *y*(*k*) не повинні бути корельовані між собою.

**Вибір кращої моделі з множини отриманих кандидатів.** На цьому етапі вибирають кращу лінійну або псевдолінійну модель за допомогою безлічі статистичних параметрів. Вони дозволяють оцінити по окремості значимість коефіцієнтів математичної моделі в статистичному змісті, визначити інтегральну помилку моделі стосовно вихідного часового ряду, встановити наявність кореляції між значеннями помилки моделі (нагадаємо, що вони повинні бути некорельованими), а також визначити ступінь адекватності моделі фізичному процесу в цілому. У цю множину входять наступні статистичні параметри:

1. *t-статистика Стьюдента*. Значимість кожного коефіцієнта регресії в статистичному змісті визначають за допомогою *t-статистики*, що, як правило, обчислюється всіма пакетами статистичних програм за формулою:

$$t_a = \frac{\hat{a} - a_0}{SE_a}$$

де $\hat{a}$ – оцінка коефіцієнта, отримана за допомогою пакета; $a_0$ – нуль-гіпотеза у відношенні значення цього коефіцієнта (звичайно $a_0 = 0$); $SE_a$ – стандартна помилка оцінки коефіцієнта, що обчислюється пакетом. Очевидно, що менше значення стандартної помилки, тим кращої, є оцінка коефіцієнта для моделі.

Для визначення значимості коефіцієнта необхідно знати довжину вибірки *N*, число оцінюваних параметрів *p* і задатися рівнем значимості α (звичайно задаються α = 1%, α = 5% або α = 10%). Рівень значимості, рівний 5%, означає, що при оцінюванні регресії ми допускаємо,

що помилкове ухвалення рішення по значимості оцінок можливо в 5% випадків. Ці параметри дозволяють вибрати по таблицях значення $t_{крит}$. Якщо

$$-t_{крит} < t_a < t_{крит},$$

то нуль-гіпотеза по незначимості коефіцієнта приймається; в іншому випадку вона відкидається і коефіцієнт вважається значимим. Оскільки значення статистики $t_a$ зворотнопропорційно стандартній помилці $SE_a$, то чим більшим буде значення $t_a$, тим більше високою буде значимість конкретного коефіцієнта.

2. *Коефіцієнт детермінації* $R^2$. Як міру інформативності часового ряду часто використовують його дисперсію. Коефіцієнт $R^2$ – це відношення дисперсії тієї частини часового ряду основної змінної, котра описується отриманим рівнянням, до вибіркової дисперсії цієї змінної. Він обчислюється за формулою:

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)}$$

Очевидно, що для адекватної моделі коефіцієнт детермінації повинен прагнути до одиниці, тобто $R^2 \to 1$.

3. *Сума квадратів помилок моделі* $\sum e^2(k)$, тобто

$$SSE = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2$$

де $\hat{y}(k) = \hat{a}_0 + \hat{a}_1 \hat{y}(k-1) + \hat{a}_2 \hat{y}(k-2) + \hat{b}_1 x(k) + \hat{b}_2 z(k)$; $y(k)$ – вимір; $N$ – довжина вибірки. Очевидно, що з можливих кандидатів необхідно вибирати ту модель, для якої $\sum e^2(k)$ приймає мінімальне значення.

4. *Інформаційний критерій Акайке* (AIC)**.** Цей критерій враховує суму квадратів помилок, число вимірів $N$ і число оцінюваних параметрів $p$:

$$AIC = N \ln\left[\sum_{k=1}^{N} e^2(k)\right] + 2p,$$

де $p$ – число оцінених параметрів. Очевидно, що для кращої моделі критерій має менше значення, оскільки він залежить від суми квадратів помилок (СКП). Однак, крім СКП, даний критерій враховує довжину вибірки і число оцінюваних параметрів, що робить його більш інформативним.

5. *Критерій Байєса-Шварца* (BSC)**.** Даний критерій схожий на попередній, однак він враховує додатково довжину вибірки за допомогою члена ln(*N*):

$$BSC = N\ln\left[\sum_{k=1}^{N} e^2(k)\right] + p\ln(N).$$

Його використовують при довгих вибірках вимірювальних даних**.**

6. *Статистика Дарбіна-Уотсона* (Durbin-Watson)

Статистика Дарбіна-Уотсона обчислюється за формулою:

$$DW = 2 - 2\rho$$

де ρ – коефіцієнт кореляції між значеннями випадкової змінної $\varepsilon(k) \approx e(k)$, тобто $\rho = \mathrm{cov}[e(k)] = E[e(k)e(k-1)]$. Цей параметр дозволяє визначити ступінь корельованості помилок моделі. При повній відсутності кореляції між помилками $DW = 2$, тобто це найбільш прийнятне значення даного параметра.

7. *Статистика Фішера F*, що визначає ступінь адекватності моделі в цілому. Для адекватної моделі виконується умова:

$$F > F_{\text{крит}},$$

де $F_{\text{крит}}$ визначається по таблиці аналогічно *t*-статистиці. Значення *F* пропорційне $R^2/(1-R^2)$, де $R^2$ – коефіцієнт детермінації. Таким чином, більшому значенню *F* відповідає більш адекватна модель.

**Приклад побудови моделі**. Розглянуту вище методику проілюструємо при побудові моделі процесу на основі вибірки даних з 120 вимірів. Для попередньої оцінки порядку авторегресійної моделі були обчислені автокореляційна і приватна автокореляційна функції. У результаті дослідження АКФ і ПАКФ встановлено наступне:

1. АКФ і ПАКФ швидко сходяться до нульових значень.

2. Теоретична АКФ процесу ковзного середнього порядку *q*, тобто СС(*q*), спадає до нуля при значенні запізнювання *q*. А теоретична АКФ процесу АР(1) спадає до нуля геометрично. У відповідності до значень АКФ процес може мати порядок 6-8, що мало відповідає дійсності.

3. Коефіцієнти ПАКФ мали такі значення: $\Phi_{1,1} = 0{,}609$; $\Phi_{2,2} = 0{,}252$. У цілому з аналізу ПАКФ можна зробити висновок, що порядок авторегресії може приймати значення 1 або 2. З іншого боку, аналіз АКФ свідчить про те, що модель може бути АР(2) або ж містити компоненти авторегресії і ковзного середнього.

4. Невеликий викид АКФ при значенні запізнювання 4 і збільшене значення ПАКФ при тому ж значенні запізнювання свідчать про те, що існує вплив вхідної змінної, затриманої на 4-му періоді дискретизації вимірів.

Зі сказаного випливає, що для математичного опису процесу необхідно скористатися моделлю АРКС(1,1) або АР(2). Можливо, знадобиться введення часу запізнювання, рівного 4. У табл. 3.7 наведені варіанти оцінювання декількох можливих структур регресійної моделі.

**Таблиця 3.7**

**Варіанти оцінювання регресійної моделі**

| | ***p*=1, *q*=0** | ***p*=2, *q*=0** | ***p*=1, *q*=1** | ***p*=1, *q*=1,4** | ***p*=1, *q*=2** |
|---|---|---|---|---|---|
| $a_0$ | 0,011<br>(4,14) | 0,011<br>(3,31) | 0,012<br>(2,63) | 0,011<br>(2,76) | 0,012<br>(2,62) |
| $a_1$ | 0,618<br>(8,54) | 0,456<br>(5,11) | 0,887<br>(14,9) | 0,791<br>(9,21) | 0,887<br>(13,2) |
| $a_2$ | | 0,258<br>(2,89) | | | |
| $\beta_1$ | | | -0,484<br>(-4,22) | -0,409<br>(-3,62) | -0,483<br>(-4,19) |
| $\beta_2$ | | | | | -0,002<br>(-0,019) |
| $\beta_4$ | | | | 0,315<br>(3,36) | |
| RSS | 0,0156 | 0,0145 | 0,0141 | 0,0134 | 0,0141 |
| AIC | -503,3 | -506,1 | -513,1 | -518,2 | -511,1 |
| BSC | -497,7 | -497,7 | -504,7 | -507,0 | -499,9 |
| *Q*(12) | 23,6(0,08) | 11,7(0,302) | 11,7(0,301) | 4,8(0,898) | 11,7(0,301) |
| *Q*(24) | 28,6(0,157) | 15,6(0,833) | 15,4(0,842) | 9,3(0,991) | 22,6(0,749) |
| *Q*(30) | 40,1(0,082) | 22,8(0,742) | 22,7(0,749) | 14,8(0,972) | 22,6(0,749) |

У дужках зазначена *t*-статистика для оцінок кожного коефіцієнта. При цьому за нульову гіпотезу прийнято, що оцінки дорівнюють нулеві. RSS (residual square sum) – сума квадратів залишків (помилок моделі). $Q(n)$ – статистика Льюнга-Бокса для автокореляції n залишків оцінюваної моделі. Для 122 вимірів основної змінної $N/4 \approx 30$. В дужках наведено рівень значимості.

Аналіз отриманих результатів дозволяє зробити наступні висновки [3]:

1. Оцінка моделі АР(3.41) підтверджує результати попереднього аналізу. Статистика Люнга-Бокса для 12 затриманих значень залишків має значення 23,6, а тому можна відхилити нуль-гіпотезу, що $Q = 0$ на рівні значимості 1%. Це свідчить про присутність істотної послідовної кореляції між помилками моделі. Таким чином, модель АР(3.41) не може бути використана для математичного опису використаного часового ряду.

2. З таблиці видно, що модель АР(2) має кращі статистичні характеристики в порівнянні з моделлю АР(1). Оцінки коефіцієнтів моделі ($\hat{a}_1 = 0{,}456$, $\hat{a}_2 = 0{,}258$) істотно відрізняються від нуля на рівні 1%, а корені характеристичного рівняння знаходяться усередині окружності одиничного радіуса. Значення $Q$-статистики свідчить про те, що автокореляція між помилками є статистично несуттєвою, тобто, нуль-гіпотеза $Q = 0$ підтверджується. Критерій AIC має менше значення для моделі АР(2). У цілому модель АР(2) краще апроксимує ряд, ніж АР(1).

3. Модель АРКС(1,1) має кращі статистичні показники, ніж АР(2). Значення *t*-статистики для оцінок коефіцієнтів (14,9 і –4,22) свідчать про високу якість оцінок. Оцінка $\hat{a}_1 = 0{,}887$ позитивна і близька до одиниці, а $Q$-статистика свідчить, що автокореляція залишків не має статистичної значимості. Критерії AIC і BSC також показують більш високу якість моделі АРКС(1,1) у порівнянні з АР(2).

4. Для того, щоб виявити присутність запізнювання на 4 періоди дискретизації, у спробну модель ковзної середньої введено додатковий член із затримкою 4. Тобто спробна модель мала вигляд

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k) + \beta_1 \varepsilon(k-1) + \beta_4 \varepsilon(k-4)$$

Відзначимо, що саме член $\beta_4\varepsilon(k-4)$ краще описує ефект запізнювання (при його наявності), ніж авторегресійний член $a_4y(k-4)$. Член ковзного середнього точніше описує такі ефекти, ніж авторегресійний. Усі коефіцієнти моделі АРКС(1,(1,4)) мають значну статистичну значимість із *t*-статистиками, рівними 9,21; –3,62 і 3,36, відповідно. Усі значення $Q$-статистики досить незначні, що свідчить про те, що автокореляція залишків статистично близька до нуля. Критерії AIC і BSC також підтримують переваги моделі АРКС(1,(1,4)).

5. Для коефіцієнта $\beta_2$ в останній розглянутій пробній моделі АРК (1,2) *t*-статистика має досить низьке значення, що дає підстави для виключення цієї моделі з подальшого розгляду.

Наступним кроком дослідження даного процесу може бути тестування часового ряду на гетероскедастичність, тобто визначення стаціонарності дисперсії ряду.