Міністерство освіти і науки, молоді та спорту України
Національний технічний університет України
"Київський політехнічний інститут"

Архипов О.Є., Архіпова С.А.

АНАЛІЗ ДАНИХ
ТА СТАТИСТИЧНА ОБРОБКА СИГНАЛІВ

Методичні вказівки до комп'ютерного практикуму

для студентів Фізико-технічного інституту НТУУ "КПІ"

спеціальності: 8.080201 – прикладна математика

Київ

НТУУ"КПІ"

2012

Архипов О.Є., Архіпова С.А. Аналіз даних та статистична обробка сигналів: методичні вказівки до комп'ютерного практикуму для студентів Фізико-технічного інституту НТУУ «КПІ» спеціальності 8.080201 "Прикладна математика"/ О.Є.Архипов, С.А.Архіпова. – К.: НТУУ«КПІ», 2012. – 82 с.

Гриф надано Вченою радою Фізико-технічного інституту НТУУ «КПІ»
(Протокол № _____ від "___" _____ р.)

Навчальне видання

АНАЛІЗ ДАНИХ ТА СТАТИСТИЧНА ОБРОБКА СИГНАЛІВ

Методичні вказівки до комп'ютерного практикуму

для студентів Фізико-технічного інституту НТУУ "КПІ"

спеціальності: 8.080201 – прикладна математика

Автори: Архипов Олександр Євгенійович, докт. техн. наук, професор

Архіпова Софія Анатоліївна, канд. техн. наук, доцент

Відповідальний редактор: Т.В.Литвинова, к. т. н., доцент

Рецензент: М.М.Савчук, д.ф.-мат. н., професор

Зміст

ВСТУП **5**

ЧАСТИНА 1. СТРУКТУРНО-ПАРАМЕТРИЧНА ІДЕНТИФІКАЦІЯ МАТЕМАТИЧНИХ МОДЕЛЕЙ **6**

1. Загальні відомості про моделі та проблему ідентифікації **6**

2. Застосування елементів регресійного аналізу в ідентифікації моделей **9**

2.1. Загальні відомості 9

2.2. Множинна лінійна регресія 12

2.3. Нелінійні регресійні моделі 14

3. Структурна ідентифікація лінійних регресійних моделей методом крокової регресії **15**

4. Параметрична ідентифікація лінійних регресійних моделей **18**

4.1. Метод найменших квадратів 18

4.2. Властивості оцінок параметрів 20

4.3. Властивості методу найменших квадратів 21

4.4. Розрахунок параметрів моделі засобами Excel 22

5. Верифікація моделей **24**

5.1. Статистика Фішера 24

5.2. Коефіцієнт детермінації 26

5.3. Статистика Стьюдента 28

5.4. Перевірка якості моделі 29

5.5. Мультиколінеарність 30

ЧАСТИНА 2. АНАЛІЗ ЧАСОВИХ РЯДІВ **33**

6. Прогнозування. Загальні аспекти. **33**

6.1. Класифікація методів прогнозування 33

6.2. Часові ряди. Структура, моделі. 36

7. Прогноз трендового компонента часового ряду **38**

7.1. Поліноміальні трендові моделі. Оцінювання якості, критерій Дарбіна-Уотсона. 38

7.2. Згладжування часових рядів зваженим ковзким середнім. Критерій Аббе. 42

7.3. Оцінка точності прогнозу за ретроданими 44

7.4. Метод гармонічних вагових коефіцієнтів 45

8. Прогнозування стохастичної складової часового ряду. Метод авторегресійних моделей. **47**

8.1. Модель авторегресії: основні поняття 47

8.2. Оцінювання структури та коефіцієнтів рівняння авторегресії 49

9. Прогноз детермінованої складової часового ряду. Метод

експоненціального згладжування. **53**
9.1. Локальні моделі часового ряду 53
9.2. Експоненціальне згладжування 55
9.3. Розрахунок оцінок параметрів і прогнозних значень 56
9.4. Вибір початкових умов 57
9.5. Вибір оптимальних параметра згладжування α та степеня полінома k методом ретропрогнозу 58
10. Особливості згладжування часових рядів в присутності аномальних даних. Усунення аномальних даних **59**
10.1. Виявлення аномальних даних в часових рядах 59
10.2. Метод медіанного згладжування 60
ЧАСТИНА 3. ЛАБОРАТОРНІ РОБОТИ **62**
№1. Параметрична ідентифікація моделей **62**
№2. Аналіз даних методом парної регресії **63**
№3. Метод крокової регресії **65**
№4. Нелінійна множинна регресія **67**
№5. Виділення тренда та прогнозування часового ряду. Статистика Дарбіна-Уотсона. **68**
№6. Оцінка точності прогнозу за ретроданими **70**
№7. Згладжування часових рядів за допомогою зваженої ковзкої середньої. Застосування критеріїв Дарбіна-Уотсона та Аббе. **71**
№8. Усунення аномальних даних в часових рядах **73**
№9. Метод гармонічних вагових коефіцієнтів **75**
№10. Метод авторегресійних моделей **76**
№11. Метод експоненціального згладжування **79**
ЛІТЕРАТУРА **80**
ДОДАТОК **82**

ВСТУП

Аналіз даних – сукупність дій, що здійснюються дослідником у ході вивчення та обробки даних з метою формування уявлень про основні закономірності та характеристики явищ чи процесів, які описуються цими даними. Традиційно аналіз даних сприймався як напрям математико-статистичних досліджень, представлений комплексом методів обробки результатів спостережень за об'єктом досліджень, функціонування якого описується множиною ознак, чиї значення припускають вимірювання та фіксацію. Метою обробки було встановлення прихованих причинно-наслідкових зв'язків на множині цих ознак з подальшою реконструкцією всього механізму функціонування об'єкту. Як правило, ці дослідження базувалися на моделях та методах регресійного аналізу, в повній мірі використовуючи розвинену методологічну базу регресійного підходу, зокрема численні критерії та процедури перевірки адекватності та точності моделей.

Стрімкий розвиток сучасних інформаційних технологій супроводжується суттєвими змінами як кількісних, так і якісних характеристик даних. Виникла необхідність аналізу величезних за обсягом та дуже різнопланових за походженням, способами вимірювання та форматами представлення масивів даних, що в свою чергу обумовило зміни погляду на базові концепції аналізу даних та технології їх реалізації. Пріоритетним стало створення автоматизованих засобів обробки даних, які у своїх діях орієнтовані на процедури аналізу, застосування та виконання яких раніше було притаманним лише людині або певним біологічним спільнотам. Цей новий напрям аналізу отримав назву "інтелектуальний аналіз даних", його підходи та методи вельми різноманітні, часто мають предметно- або галузево-орієнтований характер.

Метою даних методичних вказівок є знайомство з елементами "традиційних" технологій аналізу даних, реалізованих з використанням електронних таблиць Excel. Матеріал книги поділено на дві частини. Перша містить виклад теоретичних аспектів побудови математичних моделей за експериментальними даними, включаючи питання структурно-параметричної ідентифікації та верифікації моделей і описи відповідних лабораторних робіт. Друга – теоретичні засади застосування математичних моделей для прогнозу часових рядів, зокрема опис ряду методів прогнозування та відповідних лабораторних робіт.

ЧАСТИНА 1. СТРУКТУРНО-ПАРАМЕТРИЧНА ІДЕНТИФІКАЦІЯ МАТЕМАТИЧНИХ МОДЕЛЕЙ

1. Загальні відомості про моделі та проблему ідентифікації

У загальному випадку *модель* – це штучно створений аналог реального об'єкту (оригіналу), який певною мірою відтворює структуру, властивості та характеристики оригіналу.

Існують різні класифікації моделі залежно від ознаки, покладеної в основу класифікації. Наприклад розрізняють [6] предметні, фізичні, знакові та інші моделі. Однак серед множини моделей пріоритетне місце міцно займають *математичні моделі* (ММ), що пояснюється рядом їх визначальних особливостей: компактною формою представлення інформації, практичною відсутністю витрат на реалізацію, можливістю застосування математико-аналітичних методів аналізу та оптимізації, практично необмеженим ресурсом використання засобів обчислювальної техніки для потреб моделювання.

Розглянемо типову *постановку задачі* побудови ММ.

Маємо об'єкт дослідження (ОД), зображений на рис. 1, функціонування якого може бути описане двома групами змінних: першу з них становлять так звані вхідні, незалежні змінні або фактори, упорядкована сукупність яких утворює вектор $X = [x_1, x_2, ..., x_l]^T$, другу – вихідні, залежні змінні, вектор $Y = [y_1, y_2, ..., y_m]^T$. Для подальшого спрощення задачі вважатимемо, що вихід об’єкта дослідження є скаляром, тобто $Y = y_1 = y$.

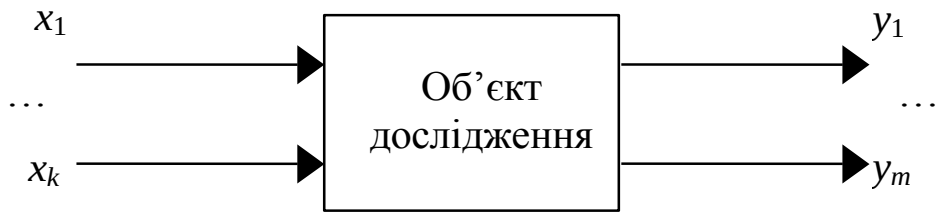

Рис. 1. Об’єкт моделювання

Тоді під ММ об'єкта досліджень будемо розуміти формальну систему, що складається з елементів $y, x_1, x_2, ..., x_l$, зв'язки між якими визначаються відношенням $\mathrm{Re}\, l(y, X)$, функція якої полягає в відтворенні

невідомого відображенні $\varphi: X \to y$, яке реалізується функціонуючим ОД. Найчастіше ММ має вигляд алгебраїчного чи диференційного рівняння, наприклад:

$$y = a_0 + a_1 x_1 + a_2 x_1 x_2 + ... + a_k \ln x_k + ... + a_p \frac{x_l}{x_1 x_2}, \qquad (1)$$

системи рівнянь, сукупності логіко-математичних співвідношень.

Слід зазначити, що не обов'язково всі вхідні змінні будуть використані у виразі (1), деякі з них можуть зовсім не впливати на залежну змінну y і ході побудови ММ мають бути вилучені з подальшого розгляду. Тобто фактори, що входять у праву частину рівняння (1), утворюють l' – вимірний вектор, для якого $l' \le l$.

Для побудови ММ зазвичай використовують інформацію, отриману шляхом спостережень за ОД у режимі його природного функціонування (пасивний експеримент) чи в ході спеціально організованих досліджень (активний експеримент). При цьому процедуру побудови ММ за експериментальними даними, отриманими при спостереженні вхідних та вихідних змінних ОД, називають *ідентифікацією об'єкта* [5] .

У загальному випадку маємо наступні *етапи ідентифікації*:

1) визначення сукупності "суттєвих" факторів, з яких формується вектор X';

2) встановлення характеру зв'язків (відношень) $\mathrm{Re}\, l(y, X')$ між елементами ММ;

3) оцінювання значень вектору $A = [a_1, a_2, ..., a_p]$ параметрів (коефіцієнтів) ММ;

4) верифікація ММ, тобто перевірка її придатності до прикладного застосування.

Сукупність елементів, що включені до складу ММ, між якими встановленні певні відношення, утворюють *структуру* Str *моделі*:

$$Str = ((y, X'), \mathrm{Re}\, l(y, X')). \qquad (2)$$

Зважаючи на це, перший і другий етапи, результатом виконання яких є визначення структури ММ, часто поєднують під назвою *структурна ідентифікація*. Безумовно структурна ідентифікація є визначальним етапом ідентифікації в цілому. Отримані на цьому етапі негативні результати унеможливлюють позитивне рішення всієї задачі. Однак інколи структура ММ може бути визначена ще до проведення експерименту з ОД шляхом безпосереднього аналізу причинно-наслідкових зв'язків, що дають змогу визначити механізми (алгоритми) дії ОД, або застосуванням сучасної системи знань про технічні, фізичні, біологічні, соціальні та інші принципи

функціонування механізмів ОД, що дозволяє виключно теоретично визначити структуру моделі. У цьому випадку немає потреби в структурній ідентифікації і процедуру ідентифікації слід починати з третього етапу – *параметричної ідентифікації* або, як її називають, ідентифікацією в вузькому сенсі [7]. Зміст задачі параметричної ідентифікації – якнайточніше обчислити кількісні значення параметрів $a_0, a_1, a_2, \ldots$ ММ за експериментальними даними, які поряд з корисною інформацією містять випадкову шумову складову, що й обумовлює появу помилок в розрахованих значеннях (оцінках) параметрів.

Останній етап ідентифікації має на меті перевірку "споживчих" якостей ММ, у першу чергу можливості її ефективного й надійного використання відповідно до мети ідентифікації. Здебільшого механізм підтвердження моделі полягає у зіставленні вихідних реакцій ОД та ММ за однією і тією самою вибіркою вхідних сигналів. Практичну однаковість цих реакцій вважають достатнім підтвердженням якості моделі.

Однак об’єктивнішими слід вважати висновки з аналізу якості ММ, отримані з урахуванням ряду аспектів, пов’язаних з цільовим призначенням ММ, яке здійснює суттєвий вплив на особливості ідентифікації та використання моделей. Зокрема, один з найпоширеніших видів ММ – це так звані *пізнавальні моделі*, що являють собою системи для накопичення, концентрації, зберігання знань та інформації навчально-пізнавального та наукового характеру. Пізнавальні моделі, не орієнтовані на застосування в конкретних прикладних задачах, тяжіють до масштабного, глобального відображення суті проблем, явищ, процесів. Прикладом таких моделей є деякі фундаментальні концепції та природні закони: модель Сонячної системи, закон всесвітнього тяжіння, моделі земної кулі та ін. Різновидом пізнавальних моделей є так звані *феноменологічні, або концептуальні моделі*, які дають спрощені схематичні описи явищ, процесів, об’єктів, уникаючи деталізації чи локальних уточнень та зосереджуючи головну увагу на найсуттєвіших елементах механізму причинно-наслідкових зв’язків у досліджуваних явищах, процесах, не враховуючи впливу менш значущих, другорядних факторів.

Інший аспект використання моделей – *прогностичні моделі*, призначені, в першу чергу, для застосування в системах управління. Цей прогноз може бути отримано як за допомогою засобів математичного моделювання, так і застосуванням моделей інших типів, наприклад, фізичних.

Серед ММ поширені так звані *апроксимативні моделі*, характерною особливістю яких є те, що структура цих моделей не відображає внутрішнього механізму тих явищ і процесів, що визначають зміст об’єкта

ідентифікації. Вибір структури апроксимативної моделі (від approximation (англ.) – наближення) реалізується тільки формально за критерієм близькості виходу моделі до значень експериментально отриманих вихідних даних за умов подання на вхід моделі реального вхідного впливу.

Сфера практичного використання апроксимативних моделей надзвичайно широка. Обставинами, що сприяють цьому, є як певні особливості ОД, так і умови проведення самого дослідження: погана структурованість та вимірюваність даних про ОД, низький рівень формалізації відомостей та інформованості про спосіб його функціонування. В подібній ситуації єдиним джерелом інформації про ОД часто є лише вибірка експериментально отриманих даних.

2. Застосування елементів регресійного аналізу в ідентифікації моделей

2.1. Загальні відомості

Побудова моделей за експериментальними даними, тобто ідентифікація ММ, має істотний теоретичний та практичний доробок, зокрема, широкий арсенал методів і засобів структурної та параметричної ідентифікації, можливості та сфера застосування яких залежить від рівня інформованості обробника щодо внутрішнього механізму дії ОД, властивостей вихідних даних, рівня випадкової похибки в цих даних тощо. Крім того, ці засоби та методи можуть мати як універсальний характер, так і бути певною мірою орієнтованими на специфічні предметні галузі, різний рівень формалізації та математизації задачі.

Зважаючи на це, в рамках загального курсу доцільно обмежитися розглядом відносно невеликої за обсягом підбірки матеріалу, зміст якого орієнтовано на певний тип ММ, який, з одного боку, достатньо розповсюджений і в своєму застосуванні не має певних галузевих вузькоспеціальних обмежень, а з іншого боку характеризується добре відпрацьованим механізмом оцінювання структури та параметрів. В достатній мірі цим вимогам відповідають так звані лінійні регресійні моделі, для структурно-параметричної ідентифікації яких застосовують регресійний аналіз. Термін "регресія" (від латинського regressio – рух назад) запропонував статистик-біолог Ф. Гальтон у 19 сторіччі. В подальшому цей термін втратив своє буквальне значення та став використовуватися для визначення залежності між взаємопов’язаними змінними. Зміна функції (залежної змінної, результату) в залежності від зміни одного чи декількох аргументів (незалежних змінних, факторів)

називається *регресією*.

Методи регресійного аналізу успішно використовуються для аналізу експериментальних даних в економіці, соціології, біології, психології, медицині, техніці тощо, причому найбільш поширені так звані лінійні регресійні моделі.

Розглянемо, чим обумовлена популярність цих моделей і чому таке звуження класу ММ під час роботи з апроксимативними моделями за певних обставин видається цілком виправданим.

Якщо припустити, що в процесі свого функціонування певний досліджуваний об'єкт описується невідомою нелінійною функцією $\varphi(x_1,...,x_l)$, яка в околі точки $X^0=[x_1^0,...,x_l^0]$ дозволяє розвинення у ряд Тейлора

$$\varphi(x_1,...,x_r)=\varphi(X^0)+\sum_{r=1}^{l}\left.\frac{\partial\varphi}{\partial x_r}\right|_{X=X^0}(x_r-x_r^0)+$$

$$+\frac{1}{2}\sum_{r=1}^{l}\sum_{j=r}^{l}\left.\frac{\partial^2\varphi}{\partial x_r\partial x_j}\right|_{X=X^0}(x_r-x_r^0)(x_j-x_j^0)+..., \qquad (3)$$

то, розкривши дужки, виконавши зведення подібних членів і замінивши в (3) усі алгебраїчні вирази, до складу яких входять лише постійні члени, коефіцієнтами a_r, $a_{r+1},...$ (наприклад $\left.\frac{\partial\varphi}{\partial x_1}\right|_{X=X^0}=a_1$), отримаємо:

$$\varphi(x_1,...,x_l)=a_0+a_1x_1+a_2x_2+...+a_lx_l+a_{l+1}x_1x_2+...+a_jx_1^2+... \quad (4)$$

Введемо в співвідношення (4) нові змінні $x_{l+1}=x_1x_2,...,\ x_j=x_1^2,...$ та обмежимо їх загальну кількість індексом p, вважаючи вплив кожної із змінних $x_{p+1},\ x_{p+2},...$ несуттєвим, а їх сукупну дію еквівалентною прояву випадкової помилки апроксимації δ. То ж маємо:

$$\varphi(x_1,...,x_l)=a_0+\sum_{k=1}^{p}a_kx_k+\delta, \qquad (5)$$

де

$$y=a_0+\sum_{k=1}^{p}a_kx_k \qquad (6)$$

– функція, що апроксимує нелінійну залежність $\varphi(x_1,...,x_l)$ в околі точки X^0.

Експериментально отримані дані про функціонування досліджуваного об'єкту звичайно містять помилки вимірювання, вплив

яких узагальнюється шляхом представлення виміряних значень $\varphi(x_1,...,x_l)$ у вигляді випадкової величини

$$z = \varphi(x_1,...,x_l) + \xi\,. \qquad (7)$$

Вводячи до виразу (7) апроксимуючу функцію (6) та застосовуючи заміну

$$e = \delta + \xi\,, \qquad (8)$$

отримуємо співвідношення

$$z = y + e = a_0 + a_1x_1 + ... + a_lx_l + a_{l+1}x_{l+1} + ... + a_jx_j + ... + a_px_p + e\,. \qquad (9)$$

Змінна e в рівнянні (9) комплексно враховує вплив випадкових збурень під час вимірювання значень $\varphi(x_1,...,x_l)$ та дію випадкового сумарного ефекту відкинутих змінних x_{p+1}, x_{p+2},... .

Якщо відносно змінних, що входять до рівняння (9), виконуються вимоги:

1) експериментально отримані значення факторних змінних $x_1,...,x_l$ не містять помилок;

2) змінна y відома лише із випадковою похибкою e , математичне очікування якої $M\{e\} = 0$, причому вплив цієї похибки від експерименту до експерименту змінюється абсолютно випадково, однак дисперсія лишається незмінною:

$$\operatorname{cov}\{e_i, e_j\} = \begin{cases} 0, & i \neq j, \\ \sigma_e^2, & i = j, \end{cases}$$

то вираз (9) відповідає *моделі класичної лінійної регресії*. Через те, що змінна *y* лінійно залежить від невідомих параметрів a_j, $j = \overline{0, p}$, регресійна модель (9) називається *лінійною за параметрами* (або просто лінійною), а співвідношення (6) – *лінійним регресійним рівнянням*. Працюючи з лінійними регресійними моделями, отримуємо суттєве спрощення розв'язку задачі ідентифікації ММ, пов'язане з тим, що задачі другого та третього етапів ідентифікації співпадають із задачами та змістом регресійного аналізу, який в своєму практичному застосуванні спирається на добре розвинений методичний та алгоритмічний апарат [9,10,11,12]. Зазначимо, що його використання обумовлює деякі термінологічні особливості стосовно формалізації та опису вирішуємих задач. Так, незалежні змінні $x_1, x_2,..., x_p$ в рівнянні лінійної регресії (9) називаються *регресорами*, параметри $a_0, a_1,..., a_p$ – *коефіцієнтами регресії*. Регресор може співпадати з факторною змінною або може бути довільною функцією від факторів, що не містить невідомих параметрів.

До змісту наведеної вище вимоги 2) в регресійному аналізі часто додається умова нормального розподілу випадкової похибки *e*, що відповідає випадку *нормальної лінійної регресії*, зокрема, рівняння (9) в цьому разі – нормальна лінійна регресійна модель (Classical Normal Linear Regression model).

Серед моделей лінійної регресії розділяють *парну регресію,* що характеризує взаємозв'язок одного результативного признаку *Y* з одним фактором *X*, та *множинну* (або *багатофакторну*) *регресію* – взаємозв'язок одного результативного признаку *Y* з декількома факторами $X_1, X_2, ..., X_p$.

Парні (прості) лінійні регресійні моделі [28] встановлюють лінійну залежність між двома змінними, наприклад врожайністю зернових від кількості опадів; витратами на відпустку та кількістю членів родини; обсягами реалізованої продукції та витратами на рекламу і т. ін.

В загальному випадку парна вибіркова регресійна модель має вигляд:

$$Z = a_0 + a_1 X + E, \qquad (10)$$

де Z – вектор значень залежної змінної, $Z = [z_1, z_2, ..., z_n]$;

X – вектор значень незалежної змінної, $X = [x_1, x_2, ..., x_n]$;

a_0, a_1 – невідомі параметри регресійної моделі;

E – вектор значень випадкових величин (помилок спостережень) , $E = [e_1, e_2, ..., e_n]$.

Модель (10) є лінійною регресійною моделлю. Її можна трактувати і як пряму на площині, де a_0 – перетин з віссю ординат або вільний член регресії, а a_1 – нахил (звичайно, якщо абстрагуватися від випадкової величини *E*).

2.2. Множинна лінійна регресія

Модель множинної (багатофакторної) регресії в загальному випадку має вигляд:

$$z = a_0 + a_1 x_1 + ... + a_l x_l + a_{l+1} x_{l+1} + ... + a_j x_j + ... + a_p x_p + e, \qquad (11)$$

причому, як це вже зазначалося у розділі 1, найважливішим етапом побудови прикладної моделі, призначеної для цілком конкретного застосування є визначення її фактичної структури. Для лінійної регресії, зважаючи що форма зв'язку між залежною змінною та регресорами вже апріорно відома, невирішеним є лише питання про склад регресорів у правій частині моделі (11). При розв'язанні цієї задачі регресійного аналізу доводиться шукати розумний компроміс між повнотою та точністю ММ. Очевидно, що із збільшенням кількості регресорів, введених до складу моделі, обов'язково буде зростати точність апроксимації моделлю вихідних даних. Однак справа в тому, що повнота та точність такої складної моделі

можуть виявитися ілюзорними із-за обмеженого обсягу даних, на яких перевіряються апроксимативні можливості моделі. З іншого боку, застосування занадто простих моделей може призвести до невиправдано великих помилок. Тому необхідні моделі розумного рівня складності. Як завжди в подібних випадках, компромісне рішення не може бути однозначним. Більше того, моделі, отримані різними методами, часто трудно порівнювати між собою або неможливо віддати остаточну перевагу якій-небудь з них. Тому коректніше казати не про оптимальну модель, а про кращу з множини розглянутих моделей.

На практиці підбір структури ММ проводиться здебільшого емпірично, шляхом побудови варіантів моделі і порівняння їх між собою за критерієм точності апроксимації ними вихідних даних.

Зазначимо, що в регресійному аналізі відсутній строго доведений алгоритм, який можна було б застосовувати для структурної ідентифікації моделі. Відомо кілька евристичних методів формування рівняння регресії вигляду (9), серед яких найпоширенішим можна вважати метод крокової регресії [8,9,10]. Цей метод являє собою процедуру послідовного покрокового відбору та введення у модель незалежних змінних, один із варіантів якої наведено нижче у розділі 3.

Нехай об’єкт досліджень реалізує невідоме відображення $\varphi: X \to Y$. Значення факторних змінних $X_1, X_2, ..., X_l$ і залежної змінної Y, представлених векторами: $X_r = [x_{r1}, ..., x_{rn}]^T$, $Y = [y_1, ..., y_n]^T$, $i = \overline{1,n}$, $r = \overline{1,l}$, отримані експериментально вимірюванням своїх значень. Припустимо, що шляхом перетворень вихідних наборів значень факторів згенерована скінченна послідовність наборів значень потенційних регресорів, перші l з яких співпадають з векторами $X_1, X_2, ..., X_l$, а наступні отримані у розрахунковий спосіб, зважаючи на те, що в загальному випадку регресор, як це вже зазначалось вище, являє собою довільну функцію факторів, яка не містить невідомих параметрів. Тоді вся інформація про об’єкт досліджень виявляється зосередженою в так званій розширеній матриці вихідних даних виду:

$$[Y, X] = \begin{bmatrix} y_1 & x_{11} & \dots & x_{l1} & \dots & x_{j1} & \dots \\ y_2 & x_{12} & \dots & x_{l2} & \dots & x_{j2} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ y_n & x_{1n} & \dots & x_{ln} & \dots & x_{jn} & \dots \end{bmatrix}. \qquad (12)$$

Зауважимо, що насправді у матриці (12) представлені ідеалізовані дані, бо, по-перше, в реальній ситуації вимірювання виконуються з

певними похибками. Звичайно припускають, що при вимірюванні значень факторів цими похибками, через їх малість можна знехтувати. Тому при описі вихідних даних враховують лише похибки вимірювання залежної змінної, вважаючи, що вони мають адитивний характер:

$$z_i = y_i + e_i,\ i = \overline{1,n}, \qquad (13)$$

де e_i – похибки вимірювань, що є реалізацією випадкової величини E_i з параметрами $M\{E_i\} = 0$, $D\{E_i\} = \sigma_{ei}^2$, z_i – реалізація випадкової величини Z_i, для якої $M\{Z_i\} = y_i$, $D\{Z_i\} = \sigma_{ei}^2$. Звичайно припускається гіпотеза гомоскедастичності, тобто рівності похибок: $\sigma_{e1}^2 = \sigma_{e2}^2 = ...\sigma_{ei}^2 = ... = \sigma_e^2$ та їх некорельованості: $\text{cov}\{E_i, E_j\} = 0$ при $i \neq j$.

Тому реальна розширена матриця вихідних даних матиме вигляд:

$$[Y + E, X] = [Z, X] = \begin{bmatrix} z_1 & x_{11} & \dots & x_{l1} & \dots & x_{j1} & \dots \\ z_2 & x_{12} & \dots & x_{l2} & \dots & x_{j2} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ z_n & x_{1n} & \dots & x_{ln} & \dots & x_{jn} & \dots \end{bmatrix}. \qquad (14)$$

По-друге, реальна структура факторних змінних невідома, тому в експерименті звичайно вимірюється те, що доступно вимірюванню та на думку дослідника, є доцільним для вимірювання. В зв'язку з цим експериментально отриманий набір факторів може бути неповним, повним або надмірним. Це ж саме стосується і згенерованого набору потенційних регресорів. Тому головною метою структурної ідентифікації в даному випадку є вибір з усієї множини потенційних регресорів їх певної підмножини, використання якої в рамках лінійної моделі (9) дозволяє достатньо повно відобразити головні властивості об'єкта досліджень [28].

2.3. Нелінійні регресійні моделі

Найбільш вивченою і методично відпрацьованою серед усіх багатовимірних регресійних моделей є лінійна. Нажаль далеко не всі соціально-економічні, природні процеси, явища та об'єкти технічного характеру можна моделювати за допомогою лінійних моделей. Їх вибір в першу чергу залежить від особливостей досліджуваного процесу (об'єкту або явища). Деякі процеси можна з певним наближенням, локально моделювати за допомогою лінійної багатофакторної моделі. Однак для повного опису процесу в межах, ширших за рамки локального наближення,

як правило, необхідно використовувати нелінійні регресійні залежності.

В даному підрозділі розглядаються нелінійні моделі [27], що припускають перетворення до лінійної форми, зокрема, логарифмуванням вихідної моделі. Як і в разі простої лінійної регресії, основне завдання полягає в розрахунку невідомих параметрів і в подальшому аналізі обраної моделі.

У ряді випадків нелінійні регресійні функції можуть бути зведені до простих лінійних застосуванням нелінійних випрямних перетворень. Для степеневої та експоненціальної регресії таке перетворення реалізується через логарифмування. Наприклад, для $y = \alpha x^{\beta}$, $y = \alpha e^{\beta x}$, $y = \alpha \beta^{x}$ відповідно маємо: $\ln y = \ln \alpha + \beta \ln x$, $\ln y = \ln \alpha + \beta x$, $\ln y = \ln \alpha + x \ln \beta$.

Увівши нові змінні: $y' = \ln y$, $a_0 = \ln \alpha$, $a_1 = \ln \beta$, $x' = \ln x$, приходимо до лінійних рівнянь: $y' = a_0 + \beta x'$, $y' = a_0 + \beta x$, $y' = a_0 + a_1 x$, що дає змогу розраховувати параметри методом найменших квадратів та використовувати подальший аналіз моделі, як і в разі простої лінійної регресії.

3. Структурна ідентифікація лінійних регресійних моделей методом крокової регресії

Для регресійних моделей існує декілька покрокових процедур (англ. *stepwise* – поступово, поетапно) відбору змінних: процедура послідовного приєднання, процедура приєднання-видалення та процедура послідовного видалення.

Ми детально розглянемо один із можливих способів організації розрахунків в покроковій процедурі приєднання, який називається методом крокової регресії.

Метод крокової регресії дає можливість визначити структуру функціональної залежності між вихідною та незалежними змінними [8,9,10].

Алгоритм методу крокової регресії містить таку послідовність дій.

1. Сформувати набір регресорів, які можуть впливати на значення вихідної змінної, зокрема, фактори (тобто вхідні змінні), квадрати факторів, їх добутки тощо.
2. Розрахувавши вибіркові коефіцієнти парної кореляції кожного регресора з вихідною змінною *Z*, вибрати та ввести регресор x_k, який має максимальний за модулем коефіцієнт кореляції, до складу початкового варіанта моделі – так званої моделі першого l=1 наближення $y_i^{(1)} = a_0^{(1)} + a_1^{(1)} x_{ki}$.

3. Обчислити вектор $A^{(1)} = [\tilde{a}_0^{(1)}, \tilde{a}_1^{(1)}]$ оцінок параметрів моделі першого наближення, розрахувати модельні значення $\tilde{y}_i$ та нев'язки цієї моделі: $\varepsilon_i^{(1)} = z_i - \tilde{y}_i^{(1)} = z_i - \tilde{a}_0^{(1)} - \tilde{a}_1^{(1)} x_{ki}$.

   Термін *нев'язка* (залишок, похибка) означає різницю між значенням залежної змінної z_i та її модельним значенням $\tilde{y}_i$: $\varepsilon_i = z_i - \tilde{y}_i$, $i = \overline{1,n}$.

4. Розрахувати суму квадратів нев'язок моделі: $S^{(1)} = \sum_{i=1}^{n} (\varepsilon_i^{(1)})^2$.

5. Розрахувати вибіркові коефіцієнти кореляції не введених у модель регресорів з нев'язками $\varepsilon_i^{(l)} = z_i - \tilde{y}_i^{(l)}$, $i = \overline{1,n}$, де l – номер наближення для попереднього кроку підбору структури моделі, визначити найбільш корельований (максимальний за модулем коефіцієнта кореляції) з нев'язками регресор і ввести його до складу моделі чергового (l+1)-го наближення.

6. Обчислити вектор $A^{(l+1)}$ оцінок параметрів ускладненої моделі на поточному l+1 кроці процедури підбору структури моделі $\tilde{y}^{(l+1)}$, розрахувати нев'язки $\varepsilon_i^{(l+1)} = z_i - \tilde{y}_i^{(l+1)}$ та відповідну суму квадратів нев'язок $S^{(l+1)} = \sum_{i=1}^{n} (\varepsilon_i^{(l+1)})^2$.

7. Розрахувати F-статистику Фішера для двох послідовно отриманих моделей для перевірки ефективності ускладнення моделі за формулою:

$$F = \frac{k\,(S^{(l)} - S^{(l+1)})}{S^{(l+1)}}, \qquad (15)$$

де

$S^{(l)}$ – сума квадратів нев'язок для регресійної моделі попереднього l-го кроку,

$S^{(l+1)}$ – сума квадратів нев'язок для моделі поточного $(l+1)$-го кроку,

k – ступінь вільності для моделі $(l+1)$-го кроку.

Ступінь вільності k дорівнює різниці між кількістю спостережень (або обсягом вибірки) n та кількістю параметрів m в моделі, тобто

$$k = n - m. \qquad (16)$$

В загальному випадку F-статистику розраховують за такою формулою:

$$F = \frac{k2}{k1}\frac{(S^{(l)} - S^{(l+1)})}{S^{(l+1)}}, \qquad (17)$$

де

$k1$ – кількість додатково введених регресорів у поточну (ускладнену) модель порівняно з попередньою (при ускладненні моделі на один елемент $k1=1$ і тоді з формули (23) отримуємо формулу (2)),

$k2$ – ступінь вільності для ускладненої моделі.

8. Порівняти розраховане значення F-статистики з критичним значенням $F_{кр}(p, k1, k2)$, що визначається за статистичними таблицями (див. додаток), де p – довірча ймовірність.

Якщо $F > F_{кр}$, чергове $(l+1)$ ускладнення моделі слід вважати доцільним: воно обумовило суттєве зменшення рівня нев'язок і покращило точність апроксимації моделлю вихідних даних. В цій ситуації виникає потреба в перевірці можливості подальшого ускладнення моделі. Тепер останню модель вже вважатимемо попередньою і задля її ускладнення перейдемо до виконання п.5.

Пункти 5–8 повторюються, поки покрокове ускладнення моделі є ефективним.

Якщо $F < F_{кр}$, останнє ускладнення структури моделі не ефективне, процедура крокової регресії переривається і в якості моделі для подальшої роботи використовується більш проста модель, а саме модель, отримана не на поточному, а на попередньому l-ому кроці підбору елементів структури моделі.

9. Як тільки процес покрокового ускладнення моделі буде перервано, слід виконати перевірку значущості оцінок параметрів найкращої моделі за t-критерієм Стьюдента:

$$t_j = \frac{\left|\tilde{a}_j\right|}{\tilde{\sigma}\{\tilde{a}_j\}}, \quad j = \overline{0, p}, \qquad (18)$$

де

$\tilde{a}_j$ – оцінка j-го параметра моделі;

$\tilde{\sigma}\{\tilde{a}_j\}$ – оцінка середнього квадратичного відхилення для $\tilde{a}_j$-го параметра моделі.

10. Розраховані значення t_j порівняти з критичними $t_{кр}(p, k)$ із статистичних таблиць t-розподілу статистики Стьюдента (див.

додаток), де p – довірча ймовірність, зазвичай p=0,95, k – ступінь вільності для обраної моделі

Для значущих параметрів моделі обов'язкова умова $t_j > t_{кр}$. Це означає, що задіяна у моделі сукупність регресорів не містить надмірності, зокрема, відсутні регресори, лінійно залежні від інших.

Якщо будуть виявлені незначущі оцінки, для яких $t_j < t_{кр}$, треба по одному виключити кожен з регресорів з моделі та виконати для отриманих таким чином спрощених варіантів моделі перевірку значущості оцінок їх коефіцієнтів за критерієм Стьюдента. Серед спрощених варіантів, що успішно пройшли перевірку, визначити найкращий (шляхом проведення зіставлення та додаткового аналізу цих варіантів за значеннями їх коефіцієнтів детермінації R^2, залишковими сумами квадратів нев'язок, *t*-статистиками), який надалі слід використовувати у якості прикладної математичної моделі.

Слід зазначити, що на практиці набагато більш поширеним способом вирішення проблеми надлишковості регресорів є видалення з моделі тих регресорів, яким відповідають незначущі оцінки коефіцієнтів. Цей спосіб не є коректним з методичної точки зору, однак у багатьох випадках він дозволяє доволі легко, а головне, оперативно отримати відносно пристойний розв'язок проблеми.

4. Параметрична ідентифікація лінійних регресійних моделей

4.1. Метод найменших квадратів

За достатнього обсягу апріорної інформації про ОД та про властивості експериментально отриманих вихідних даних, зокрема, про характеристики їх шумової складової *e*, можливе застосування широкого кола добре відпрацьованих методів і способів структурної та параметричної ідентифікації [5,7,12]. Серед останніх найвідомішим є метод максимальної правдоподібності [10,12], який у ряді випадків, спираючись на відомі статистичні характеристики вихідних даних, дозволяє побудувати ефективні алгоритми параметричної ідентифікації.

Однак при побудові моделі ОД за експериментально отриманими даними поширеною є ситуація, для якої практично вся інформація, що використовується обробником для розв'язання поставленої задачі, обмежується вибіркою вихідних даних. Тому для розв'язання задачі

параметричної ідентифікації використовують методи, орієнтовані виключно на інформацію про нев'язки моделі.

В загальному випадку для довільної моделі $y = f(X, A)$ відомої структури рівень нев'язок $\varepsilon_i = z_i - \tilde{y}_i$, $i = \overline{1,n}$ залежить тільки від вибору параметрів моделі, тобто елементів вектора $A = [a_0, a_1, ..., a_l]$. Тому, якщо ввести показник якості параметричної ідентифікації, який інтегрує в собі всю інформацію про рівні нев'язок ε_i, $i = \overline{1,n}$ і містить відомості про залежність рівня нев'язок від значень параметрів моделі, то екстремізація (зазвичай мінімізація) цього показника дозволить визначити оптимальні параметри моделі.

Найпоширенішими в задачах параметричної ідентифікації є два типи показників:

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (z_i - f(X_i, A))^2 \qquad (19)$$

та

$$Q_1 = \sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |z_i - f(X_i, A)|. \qquad (20)$$

Оскільки структура $f(X_i, A)$ не змінюється під час оцінювання параметрів $A = [a_0, a_1, ..., a_l]$, кількісні значення функціоналів Q, Q_1 є залежними тільки від вибору значень оцінок $a_0, a_1, ..., a_l$, тобто Q, Q_1 можна розглядати як функції цих параметрів: $Q = Q(a_0, a_1, ..., a_l)$, $Q_1 = Q_1(a_0, a_1, ..., a_l)$. Тоді задачу параметричної ідентифікації можна сформулювати так: підібрати на множині $\{A\}$ можливих значень параметрів моделі такі значення оцінок $\tilde{A}$, щоб функції $Q(A)$, $Q_1(A)$ досягли своїх мінімумів, тобто метою цього аналізу є пошук

$$\tilde{A} = \arg \min_{A \in \{A\}} Q(A), \quad \tilde{A}_1 = \arg \min_{A \in \{A\}} Q_1(A). \qquad (21)$$

На практиці більш поширеним є показник Q, бо функція $Q(A)$ неперервно залежить від $a_0, a_1, ..., a_l$ і дозволяє обчислити частинні похідні $\partial Q / \partial a_i$, $i = \overline{0,l}$, скласти з них систему рівнянь виду

$$\frac{\partial Q}{\partial a_i} = 0, \quad i = \overline{0,l} \qquad (22)$$

та знайти оптимальні параметри моделі. Отримані таким шляхом оцінки називають *МНК-оцінками* (оцінками за методом найменших квадратів), бо оптимізація параметрів моделі призводить до мінімуму (19) показника Q, тобто за мінімумом суми квадратів нев'язок.

Оцінки, знайдені за умов мінімуму (20) – мінімуму суми модулів нев'язок, називають *МНМ-оцінками* (оцінками за методом найменших модулів).

При оцінюванні параметрів регресійної моделі, виходячи з умов мінімізації суми квадратів нев'язки, загальні формули для обчислення МНК-оцінок легко отримати у матричній формі запису:

$$Q = \varepsilon^T \varepsilon = (Z - X\tilde{A})^T (Z - X\tilde{A}) = Z^T Z - 2Z^T X\tilde{A} + \tilde{A}^T X^T X\tilde{A},$$
$$\frac{\partial Q}{\partial \tilde{A}} = -2X^T Z + 2X^T X\tilde{A} = 0, \qquad (23)$$
$$\tilde{A} = (X^T X)^{-1} X^T Z.$$

Матриця коваріацій МНК-оцінок [9,10,11] має вигляд

$$C\{\tilde{A}\} = \sigma_e^2 (X^T X)^{-1}. \qquad (24)$$

Оцінка дисперсії помилки розраховується за формулою:

$$\tilde{\sigma}_e^2 = \frac{1}{n-l} \sum_{i=1}^{n} \varepsilon_i{}^2 = \frac{S^{(l)}}{n-l}, \qquad (25)$$

де l – кількість параметрів регресійної моделі, $S^{(l)}$ – сума квадратів нев'язок цієї моделі.

4.2. Властивості оцінок параметрів

Виникають запитання: наскільки задовільні отримані вище МНК-оцінки, чи можна їх поліпшити та наскільки це доцільно, наскільки вдалим є вибраний спосіб обчислення оцінок і, нарешті, наскільки адекватною, прийнятною є обрана структура моделі. Відповіді на ці запитання тією чи іншою мірою дає математична статистика [2,3,9–13].

Уведемо деякі визначення.

Сукупність значень, отриманих у ході спостереження ряду величин під час проведення досліджень (експериментів), називають *вибіркою*. Кількість спостережень (експериментів) n визначає обсяг вибірки.

Будь-яку функцію вибіркових даних, що не залежить від невідомих параметрів, називають *статистикою*. Статистика є випадковою величиною.

Статистику, обчислену за вибірковими даними, яку беруть як невідоме значення параметра, називають *оцінкою* цього параметра.

Якість оцінки, її близькість до невідомого істинного значення параметра характеризується трьома властивостями: незсуненістю, спроможністю та ефективністю.

Оцінка $\tilde{a}$ є *незсуненою*, якщо її математичне сподівання дорівнює істинному значенню a параметра:

$$M\{\tilde{a}\} = a. \qquad (26)$$

Розраховану оцінку $\tilde{a}_n$ за вибіркою обсягом n називають *спроможною*, якщо вона збігається за ймовірністю P до істинного значення a параметра, тобто для будь-якого $\delta > 0$

$$P\{|\tilde{a}_n - a| \le \delta\} \to 1 \quad \text{при } n \to \infty. \qquad (27)$$

Оцінка $\tilde{a}_{ef}$ є *ефективною* в певному класі оцінок $\tilde{A}$, якщо вона найбільш точна серед цих оцінок, тобто має найменшу дисперсію:

$$D\{\tilde{a}_{ef}\} = \min_{\tilde{a}^{(q)} \in \tilde{A}} D\{\tilde{a}^{(q)}\} \qquad (28)$$

Наприклад, якщо невідоме значення a параметра можна оцінити кількома методами параметричної ідентифікації $\pi_1,...,\pi_q,...,\pi_k$, внаслідок чого отримуємо сукупність оцінок $\tilde{A} = [\tilde{a}^{(1)},...,\tilde{a}^{(q)},...,\tilde{a}^{(k)}]$, які відповідно мають дисперсії $\sigma_1^2,...,\sigma_q^2,...,\sigma_k^2$, то ефективною буде та з оцінок, що має мінімально можливу дисперсію.

Аналіз властивостей МНК-оцінок без припущення нормальності розподілу помилок дозволяє стверджувати [10] єдинність, незсуненість і ефективність оцінок (23) у класі лінійних незсунених оцінок, а за достатньо слабких додаткових обмежень – також спроможність та асимптотичну нормальність. Оцінка дисперсії (25) є незсуненою і спроможною.

У випадку нормальності розподілу помилки E МНК-оцінки параметрів регресії ефективні в класі усіх незсунених оцінок (як лінійних, так і нелінійних) і мають нормальний розподіл із середнім A та коваріаційною матрицею $C\{\tilde{A}\}$ [10].

4.3. Властивості методу найменших квадратів

1. Сума вихідних даних дорівнює сумі значень, розрахованих за моделлю:

$$\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} \tilde{y}_i \cdot \qquad (29)$$

2.Сума нев'язок дорівнює нулю:

$$\sum_{i=1}^{n} \varepsilon_i = \bar{\varepsilon} = 0 \cdot \qquad (30)$$

3. Нев'язки ε_i некорельовані з $x_{1i}, x_{2i},..., x_{pi}$, тобто

$$\sum_{i=1}^{n}\varepsilon_i x_{1i} = \sum_{i=1}^{n}\varepsilon_i x_{2i} = ..., = \sum_{i=1}^{n}\varepsilon_i x_{pi} = 0 \cdot \quad (31)$$

4. Нев'язки ε_i некорельовані з $\tilde{y}_i$, тобто

$$\sum_{i=1}^{n}\varepsilon_i \tilde{y}_i = 0 \cdot \quad (32)$$

5. Оцінки, розраховані за МНК, якщо виконуються усі припущення щодо випадкової величини *E*, є лінійними, без відхилень, мають найменшу дисперсію з усіх можливих методів оцінювання.

Таким чином, метод найменших квадратів є найкращим методом для оцінювання невідомих параметрів лінійної регресії.

Властивості методу найменших квадратів у випадку простої лінійної моделі збігаються з випадком багатофакторної регресії.

4.4. Розрахунок параметрів моделі засобами Excel

Регресійний аналіз передбачає значний об'єм необхідних обрахунків та застосування складних рівнянь для аналізу великої кількості даних. З появою потужної обчислювальної техніки використання методів регресійного аналізу стало більш доступним та успішно застосовується для аналізу експериментальних даних в економіці, соціології, біології, психології, медицині, техніці тощо.

Електронні таблиці Microsoft Excel пропонують широкий діапазон засобів для аналізу та обробки даних [24–26]. Головна особливість програми Excel – табличне подання даних, яке є типовим для відображення інформації різноманітного призначення: економічної, соціальної, виробничої або персональної.

При побудові математичних моделей з використанням електронних таблиць, по-перше, не губиться алгоритм розв'язку задачі; по-друге, обробник звільняється від рутинної роботи розрахунків і, по-третє, студент навчається досконало володіти електронними таблицями і використовувати комп'ютерні технології для вирішення практичних задач.

Оцінка параметрів моделі за допомогою функції ЛИНЕЙН

При виконанні лабораторного практикуму частіше використовуються вбудовані функції Excel статистичної, математичної категорій та Пакету аналізу. Зокрема, якщо математична модель описується рівнянням багатофакторної лінійної регресії $Y = a_0 + a_1 X_1 + a_2 X_2 + ... + a_p X_p$, для визначення оцінок параметрів $\tilde{A} = [\tilde{a}_0, \tilde{a}_1, ..., \tilde{a}_p]$, $j = \overline{0, p}$ за методом найменших квадратів використовують в Excel статистичну функцію ЛИНЕЙН з чотирма аргументами:

ЛИНЕЙН (Z; X ; *константа; статистика*),

де Z – вектор експериментально отриманих значень залежної змінної *y*, $z_i = y_i + e_i$, e_i – помилка вимірювання *i*-тої залежної змінної, $i = \overline{1,n}$, n – об'єм вибірки;

X – матриця значень незалежних змінних, де кожна змінна $X_1, X_2, ..., X_p$ є вектор-стовпець, тобто

$$Z = \begin{bmatrix} z_1 \\ z_2 \\ ... \\ z_n \end{bmatrix}; \qquad X = [X_1, X_2, ..., X_p] = \begin{bmatrix} x_{11} & x_{21} & ... & x_{p1} \\ x_{12} & x_{22} & ... & x_{p2} \\ ... & ... & ... & ... \\ x_{1n} & x_{2n} & ... & x_{pn} \end{bmatrix};$$

константа – дорівнює 1, якщо в моделі є вільний член a_0, у противному разі константа дорівнює 0;

статистика – приймає два значення 0 або 1, яке вказує, чи потрібно повертати додаткову статистику по регресії: якщо "статистика" дорівнює 0, то функція ЛИНЕЙН повертає тільки рядок параметрів a_j, $j = \overline{0,p}$; якщо "статистика" дорівнює 1, то функція поверне додаткову регресійну статистику, яку розташує у п'яти строках.

Для роботи з функцією треба передчасно виділити в робочому листі Excel блок комірок розміром: кількість стовпців дорівнює кількості оцінок невідомих параметрів $\tilde{a}_0, \tilde{a}_1, ..., \tilde{a}_p$, а кількість рядків дорівнює п'яти, тобто розмір блоку: (p+1) стовпців на 5 рядків.

Після завершення роботи з вікном функції ЛИНЕЙН для заповнення всього блоку комірок треба натиснути функціональну клавішу <F2>, далі одночасно натиснути три клавіші <Ctrl >+ <Shift >+ <Enter>.

Схема розташування додаткової статистики, яку розраховує функція ЛИНЕЙН знаходиться у табл.1.

Таблиця 1. Додаткова регресійна статистика функції ЛИНЕЙН

$\tilde{a}_p$	$\tilde{a}_{p-1}$	...	$\tilde{a}_2$	$\tilde{a}_1$	$\tilde{a}_0$
Sa_p	Sa_{p-1}	...	Sa_2	Sa_1	Sa_0
R^2	S_Y				
F	k				
Sreg	*Sost*				

де

$\tilde{a}_0, \tilde{a}_1, ..., \tilde{a}_p$ – оцінки коефіцієнтів $a_0, a_1, ..., a_p$;

$Sa_0,...,Sa_p$ – стандартні похибки для оцінок коефіцієнтів;

R^2 – коефіцієнт детермінації;

S_Y – середнє квадратичне відхилення для оцінки залежної змінної *Y*;

F – F -статистика Фішера;

k – ступінь вільності;

$Sreg$ – регресійна сума квадратів, $Sreg = \sum_{i=1}^{n} (\tilde{y}_i - \bar{z})^2$, $i = \overline{1,n}$

де $\tilde{y}_i$ – модельні значення залежної змінної,

$\bar{z}$ – середнє арифметичне значення залежної змінної;

$Sost$ – залишкова сума квадратів нев'язок, $\varepsilon_i = z_i - \tilde{y}_i$, $Sost = \sum_{i=1}^{n} (z_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$, де z_i – значення залежної змінної.

5. Верифікація моделей

Як зазначалося у першому розділі, верифікація математичної моделі – це перевірка ефективності її застосування відповідно до вихідної мети моделювання, корисність її використання для вирішення певної прикладної задачі. Через це верифікувати модель у повному обсязі можна лише у процесі її функціонування "за місцем призначення", тоді як на етапі побудови ММ можлива тільки перевірка її якісних показників за системою певних критеріїв. Зокрема, щодо регресійних моделей, такій перевірці відповідає комплексний аналіз якості моделі за *F*-, *t*-статистиками, за значеннями коефіцієнту детермінації R^2, за величиною залишкової суми квадратів нев'язок тощо.

Нижче наведено коротке зведення основних теоретичних відомостей щодо цих статистик та способів їх практичного застосування у аналізі якості регресійних моделей.

5.1. Статистика Фішера

Існує декілька статистик Фішера, які використовуються залежно від поставленої мети дослідження.

Кількісно ступінь прийнятності гіпотези лінійної залежності між *y* та *x* можна обчислити за допомогою наступної статистики Фішера

$$F = \frac{n-m}{n-1} \frac{\sum_{i=1}^{n} (\tilde{y}_i - \bar{z})^2}{\sum_{i=1}^{n} (z_i - \tilde{y}_i)^2}, \qquad (33)$$

де m – кількість параметрів моделі, n – об'єм вибірки, $\bar{z}$ – середнє арифметичне значення залежної змінної Z. Статистика Фішера показує у скільки разів лінійна регресія краще описує взаємозв'язок $Y = f(X)$ ніж заміна її середнім $\bar{z}$.

Недоліком цієї статистики є постійне порівняння всіх моделей, від загрублених до переускладнених, з середнім залежної змінної. Зрозуміло, що всі ці моделі будуть краще ніж $\bar{z}$.

Частіше використовується F-статистика Фішера для двох послідовно отриманих моделей з яких друга утворена шляхом ускладнення першої (попередньої) моделі за рахунок введення нового регресора. Розраховане значення статистики зіставляють з критичним значенням $F_{кр}$, яке обирають зі статистичних таблиць.

Емпіричне значення F-статистики розраховується за формулою:

$$F = \frac{k\,(S^{(l)} - S^{(l+1)})}{S^{(l+1)}}, \qquad (34)$$

де $S^{(l)}$ – сума квадратів нев'язок для регресійної моделі l-го кроку,

$S^{(l+1)}$ – сума квадратів нев'язок для моделі наступного $(l+1)$-го кроку,

$k = n-m$ – ступінь вільності для моделі наступного $(l+1)$-го кроку.

Статистика Фішера F використовується для прийняття рішення про доцільність продовження ускладнення моделі, що базується на аналізі відносного приросту точності моделі після чергового ускладнення її структури. Чергове ускладнення моделі через введення до існуючої моделі додатково регресора слід вважати доцільним, якщо це призвело до суттєвого зменшення рівня $S^{(l+1)}$ порівняно із рівнем $S^{(l)}$ попередньої моделі.

За таблицями F-розподілу (див. додаток) визначити критичне значення $F_{кр}(p, k1, k2)$, де p – довірча ймовірність (зазвичай p=0,95 або p=0,99), $k1$ – кількість введених параметрів у нову модель (при ускладненні моделі на один елемент $k1 = 1$), $k2$ – ступінь вільності для складної моделі.

Для кількісної підтримки рішення про доцільність ускладнення моделі необхідно зіставити значення F та $F_{кр}$.

Якщо $F \geq F_{кр}$, ускладнення моделі слід вважати доцільним і можна продовжити процес ускладнення моделі за рахунок включення до неї

нового регресора. При цьому ускладнена модель на попередньому етапі набуває статусу початкової моделі й починається черговий крок побудови наступної регресійної моделі.

Якщо $F < F_{кр}$, це означає, що ускладнення моделі за рахунок введення чергового регресора практично не додало точності моделі: різниця $S^{(l)} - S^{(l+1)}$ порівняно із значенням $S^{(l+1)}$ є дуже малою, процес ускладнення моделі стає неефективним. Тому можна перервати процедуру крокової регресії, узявши за остаточний, базовий варіант модель, отриману на попередньому кроці, тобто до введення останнього регресора. Цю модель можна вважати найкращою із побудованих.

Досвід застосування *F*-критерію Фішера показав [13], що остаточне рішення про визначення структури моделі слід робити лише в тому разі, коли два послідовних ускладнення моделі будуть обидва неефективними. Наразі, коли друге ускладнення моделі виявилось ефективним, тобто $F \geq F_{кр}$, слід продовжити процедуру крокової регресії для визначення структури моделі.

Наведена вище схема алгоритму крокової регресії не є єдино можливою, а сам метод крокової регресії не гарантує стабільно кращих розв'язків задачі структурної ідентифікації, хоча й набув за твердженням [19] найбільшого практичного поширення. Однією з вад побудованих цим методом моделей є їх надлишковість. Зокрема скінчений обсяг *n* вихідних даних обумовлює можливість помилкового включення до структури моделі зайвих регресорів, для яких у випадку $n \to \infty$ значення відповідних коефіцієнтів моделі мали б дорівнювати 0. Виявлення подібних "підозрілих" регресорів здійснюється за критерієм Стьюдента й розглянуто у розділі 5.3.

5.2. Коефіцієнт детермінації

Коефіцієнт детермінації R^2 визначає апроксимативні якості побудованої регресійної моделі та перевіряє її адекватність (відповідність) реальним даним.:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (z_i - \tilde{y}_i)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2} \,. \qquad (35)$$

Значення коефіцієнта детермінації лежать в інтервалі $0 \leq R^2 \leq 1$. Якщо $R^2 = 0$, то рівняння регресії невдало описує вихідні дані і за

точністю їх апроксимації співпадає з тривіальною моделлю $\tilde{Y} = \bar{z}$. Крайній випадок $R^2 = 1$ визначає наявність точного лінійного зв'язку між Z та X, тобто всі реальні вихідні дані співпадають з модельними. Якщо взяти число регресорів рівним числу спостережень, завжди можна домогтися того, що $R^2 = 1$, але це зовсім не буде означати наявність змістовної моделі.

Важливою властивістю коефіцієнта детермінації R^2 є те, що він – не спадна функція від кількості факторів, які входять до моделі. Якщо кількість факторів зростає, то R^2 також зростає і ніколи не зменшується. У виразі (35) знаменник не залежить від кількості факторів x, тоді як чисельник, який дорівнює сумі квадратів нев'язок, навпаки, залежить: із збільшенням кількості факторів x величина суми квадратів нев'язок спадає.

Тому при виборі кращої моделі не слід покладатися тільки на коефіцієнт детермінації.

Для усунення ефекту зростання R^2 при ускладненні моделі розраховується *скоригований коефіцієнт детермінації*

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-m}\sum_{i=1}^{n}(z_i - \tilde{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2}. \qquad (36)$$

Завдяки корекції в рівнянні (36) маємо незміщену оцінку дисперсії помилок в чисельнику, а в знаменнику – незміщену оцінку дисперсії значень величини Z.

У певній мірі використання скоригованого коефіцієнта детермінації $\bar{R}^2$ більш коректне для порівняння моделей при зміні кількості регресорів.

Середньо-квадратичне відхилення (СКВ) для оцінки залежної змінної Z розраховують за формулою:

$$\sigma = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}(z_i - \tilde{y}_i)^2}. \qquad (37)$$

5.3. Статистика Стьюдента

Для базової моделі треба перевірити значимість отриманих оцінок $\tilde{a}_0, \tilde{a}_1, ..., \tilde{a}_p$, для чого для кожної оцінки розраховують *t*-статистики Стьюдента за формулою:

$$t_j = \frac{|\tilde{a}_j|}{\tilde{\sigma}\{\tilde{a}_j\}}, \quad j = \overline{0, p}, \tag{38}$$

де $\tilde{\sigma}\{\tilde{a}_j\}$ – оцінка середнього квадратичного відхилення для $\tilde{a}_j$-го параметра моделі. Розраховані статистики порівнюються з критичними значеннями $t_{кр}(k, p)$, де k – ступінь вільності для заданої довірчої ймовірності p, які визначаються за таблицями (див. додаток) *t*-розподілу статистики Стьюдента.

Для значущих оцінок параметрів буде виконуватися умова $t_j > t_{кр}$. Якщо для оцінки $\tilde{a}_j$ маємо $t_j < t_{кр}$, можна стверджувати, що з ймовірністю p оцінка $\tilde{a}_j$ є статистично незначущою й, можливо, відповідний регресор x_j слід вилучити зі структури моделі.

Фактично у наведеній перевірці досліджується рівень відмінностей відповідних оцінок $\tilde{a}_j$ від 0, тобто перевіряється нуль-гіпотеза H_0: $a_j = 0$ проти гіпотези H_1: $a_j \neq 0$ (слід однак мати на увазі можливість існування альтернативної гіпотези, пов'язаної з проблемою мультиколінеарності набору регресорів, про яку більш докладно йдеться у підрозділі 5.5). За спрощеною процедурою, що не потребує залучення таблиці *t*-розподілу для перевірки гіпотез, нуль-гіпотеза приймається при $t_j < 2$ [3], тобто параметр a_j є незначущий.

Якщо при вилученні регресора фіксується незначне зменшення значення коефіцієнту детермінації R^2, тобто модель, не зважаючи на спрощення, майже не втрачає у точності, виключений регресор не має сенсу повертати до структури моделі. Більш-менш істотне зменшення значення R^2 ставить під сумнів необхідність виключення "підозрілого" регресора. В цьому випадку необхідно визначити значимість інших параметрів моделі, зокрема параметру a_0, необхідність введення якого до моделі ніяк не перевірялася.

5.4. Перевірка якості моделі

Застосовуючи апарат регресійного аналізу для розв'язання задач структурно-параметричної ідентифікації моделей, необхідно мати на увазі, що отримані статистично значимі висновки не мають жорстко імперативного характеру, це умовно-допоміжна інформація, яку повинен враховувати обробник, приймаючи, наприклад, рішення про вилучення, залишення чи включення до моделі того чи іншого регресора. Кожне з подібних рішень потребує додаткової перевірки чи підтвердження за іншими статистичними показниками. Так, аналізуючи конкуруючі варіанти моделей (зокрема, зіставляючи початкову та ускладнену моделі), слід брати до уваги комплекс показників, що характеризує точність та адекватність моделей даним. Це коефіцієнт детермінації R^2, сума квадратів помилки апроксимації S_l, F-статистика Фішера тощо.

Перевірка якості моделі частіше за все зводиться до аналізу покращення точності моделі на черговому етапі її ускладнення порівняно з точністю моделей, отриманих на попередніх кроках, тобто порівнюється кількісне значення суми квадратів нев'язок $S^{(l+1)}$ із показниками $S^{(l)}, S^{(l-1)}, ..., S^{(1)}$.

Для повноти картини доцільно ввести нульове l=0 наближення $\tilde{y}^{(0)} = \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$, що відповідає найпростішій моделі середнього арифметичного залежної змінної. Для цієї моделі показник точності обчислюється за формулою $S^{(0)} = \sum_{i=1}^{n}(z_i - \bar{z})^2$.

Динаміка зміни показника $S^{(l)}$ представлена на рис. 2.

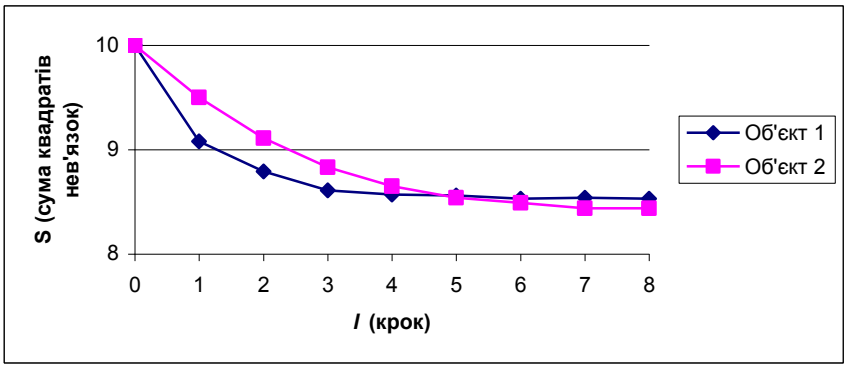

Рис. 2. Залежність суми квадратів нев'язки $S^{(l)}$ від кроку ускладнення моделі *l*.

Графіки на рис. 2 показують, що підбір моделі для Об'єкту 1 можна скінчити вже на третьому кроці, бо подальше ускладнення структури моделі ніяк не сприяє покращенню її точності, значення показників $S_3, S_4, S_5, \ldots$ практично не змінюються. Для Об'єкту 2 ускладнення структури моделі можна припинити на 5-7 кроці, залежно від вимог, що висуваються до якості моделі. Наприклад, якщо якість моделі розуміти як компроміс між критеріями простоти та точності, слід, припинити ускладнення моделі вже на 5-му кроці. Орієнтуючись тільки на вимоги точності, процедуру ускладнення моделі можна подовжити до 6-7 кроків.

Аналогічні міркування можна застосувати до аналізу динаміки коефіцієнту детермінації R^2, додатково використовуючи властивість коефіцієнту детермінації асимптотично наближатися до 1 при ускладненні моделі.

Таким чином, необхідно аналізувати динаміку змін значень усіх показників у процесі підбору моделі. Зокрема, за деякими літературними джерелами рекомендується не переривати процедуру крокової регресії відразу після першого висновку про недоцільність введення чергового регресора до структури моделі, а лише після підтвердження цієї негативної тенденції впродовж одного-двох наступних кроків.

5.5. Мультиколінеарність

Однією з умов класичної лінійної регресії є припущення про лінійну незалежність регресорів, що на практиці означає лінійну незалежність стовпчиків (векторів) матриці регресорів [*X*] . Якщо це припущення не виконується, зокрема хоча б один з векторів можна представити лінійною комбінацією інших, кажуть, що має місце так звана *повна мультиколінеарність* (мультиколінеарність в строгому розумінні, досконала, сильна, строга) між регресорами. У регресійній моделі з *p* регресорами $x_1, x_2, \ldots, x_p$ існуванню повної мультиколінеарності відповідає виконання такої умови:

$$b_1 x_1 + b_2 x_2 + \ldots + b_p x_p = 0,$$

де не всі коефіцієнти b_j, $j = \overline{1, p}$ одночасно дорівнюють нулю. А саме, якщо два чи більше регресорів мають відмінні від нуля коефіцієнти, то будь який з цих регресорів може бути представлений лінійною комбінацією інших. Зокрема, якщо таких регресорів буде лише два: X_i, X_j, вони мають задовольняти умові парної лінійної залежності, тобто $r_{x_i x_j} = 1,\ i \neq j$. В

такій ситуації фактично неможливо оцінити окремий вплив кожного з цих регресорів на досліджуваний показник Y.

Зазначимо, що для повної мультиколінеарності детермінант $\det(X^T X) = 0$, тому стає неможливим виконати розрахунок МНК-оцінок коефіцієнтів регресії.

Однак на практиці повна мультиколінеарність зустрічається вкрай рідко [2,12,26,27]. Набагато частіше зустрічається стохастична форма мультиколінеарності (або просто *мультиколінеарність*), за якої в багатофакторній регресійній моделі дві або більше незалежні змінні (регресори) пов'язані між собою майже лінійною залежністю, тобто мають високий ступінь кореляції ($r_{x_i x_j} \to 1,\ i \neq j$). Чим більш щільний кореляційний зв'язок між регресорами, тим менше $\det(X^T X)$, що призводить до значного зменшення точності оцінки $\tilde{A}$, викривленню оцінок дисперсій залишків, дисперсій параметрів регресії і коваріацій між ними.

Якщо мультиколінеарність значна, визначити МНК-оцінки коефіцієнтів регресії можна, але їхні середньоквадратичні відхилення будуть дуже великими. Як наслідок, значення параметрів для сукупності неможна визначити точно.

Значній кореляції між факторами x_j, $j = \overline{1, p}$ відповідає наступне співвідношення:

$$b_1 x_1 + b_2 x_2 + ... + b_p x_p + E = 0,$$

де E – випадкова величина, тим менша, чим вищий рівень корельованості регресорів.

Практичні наслідки мультиколінеарності:

- велика дисперсія і коваріація оцінок параметрів, обчислених за методом найменших квадратів;
- чутливість оцінок параметрів до обсягів сукупності спостережень;
- незначущість *t*-статистик Стьюдента, тобто параметри багатофакторної регресії прямують до нуля.

Хоча надійних методів тестування мультиколінеарності не існує, є декілька її *ознак*.

- незначні зміни у вихідних даних (наприклад, додавання нових даних) призводять до істотних змін оцінок коефіцієнтів регресії;
- високе значення коефіцієнту детермінації при незначущості параметрів за *t*-статистикою Стьюдента;

- у моделі з двома регресорами – велике значення парного коефіцієнта кореляції між цими регресорами;
- оцінки коефіцієнтів регресії стають невиправдано великими за своїми значеннями.

Усі ці ознаки мультиколінеарності мають один спільний недолік: жодна з них чітко не розмежовує випадки, коли мультиколінеарність істотна і коли нею можна знехтувати.

Виявлення мультиколінеарності є лише першою частиною справи. Друга частина – як *позбутися* мультиколінеарності. В залежності від особливостей задачі обробник може застосувати алгоритм Фаррара-Глобера, метод головних компонентів, метод усунення змінної з мультиколінеарної пари.

На жаль, немає якихось універсальних методів, але для усунення мультиколінеарності можна:

- використати додаткову або первинну інформацію;
- об'єднати інформацію;
- відкинути змінну з високою кореляцією;
- перетворити дані;
- збільшити обсяг спостережень.

Що спрацює на практиці, залежить від істотності проблеми та її характеру.

В умовах обмеженого обсягу даних та недостатньої інформації про досліджуваний об'єкт для запобігання мультиколінеарності вхідних змінних треба виділити пари (x_i, x_j), для яких коефіцієнти парних кореляцій $r_{x_i x_j} > 0{,}9$ і далі у формуванні структури регресійної залежності використовувати лише одну змінну з кожної пари. Ця змінна має мати більший коефіцієнт кореляції із залежною змінною y, наприклад, якщо $r_{yx_i} > r_{yx_j}$, слід вибрати змінну x_i.

ЧАСТИНА 2. АНАЛІЗ ЧАСОВИХ РЯДІВ

6. Прогнозування. Загальні аспекти.

6.1. Класифікація методів прогнозування

Під *прогнозуванням* будемо розуміти науково обґрунтоване оцінювання можливого майбутнього стану явищ і процесів, які є об'єктом аналізу, кількісних та якісних змін цього стану або показників, що його характеризують, строки та форми реалізації цих змін.

Наукова галузь, пов'язана з дослідженням і вивченням методології та закономірностей прогнозування, називається *прогностикою*.

До головних *категорій прогностики* належить:

- об'єкт прогнозування – процеси, явища, події, на які орієнтована пізнавальна та практична діяльність людини, прогнозний фон;
- сукупність зовнішніх щодо об'єкта умов, суттєвих для обґрунтування прогнозу;
- метод прогнозування – спосіб дослідження об'єкта прогнозування, спрямований на розробку прогнозу;
- методика прогнозування – сукупність спеціальних правил, прийомів і методів прогнозування.

Нині відомо понад сто методів прогнозування [1,2,4,15,16], що робить актуальною проблему їх класифікації. Є різні системи класифікації методів прогнозу. Залежно від обраних критеріїв (класифікаційних ознак) з літератури відомо близько двох десятків систем класифікації [14,15]. Чи не найбільш цікавою та актуальною є класифікація за часом (періодом, горизонтом) упередження L, тобто часова градація прогнозів за їх терміном (строком), якому відповідає відрізок часу, що відділяє момент t_n, у який було останній раз отримано дані про об’єкт прогнозування, від моменту $t_n + L$, для якого визначають прогноз. Частіше за все виділяють п’ять типів прогнозу [15]:

– поточний (оперативний) прогноз, що орієнтований на строк упередження, в межах якого можливі кількісні зміни параметрів об’єкта дослідження;

– довготерміновий (довгостроковий) прогноз – в межах його можливі якісні (структурні) зміни об’єкта дослідження та оточуючого середовища, в якому він функціонує;

– середньотерміновий (середньостроковий) прогноз – проміжний між коротко- та довготерміновим з можливою перевагою очікуваних кількісних змін над якісними;

– наддовготерміновий (наддовгостроковий) прогноз, в межах якого можливі вельми суттєві якісні (структурні) зміни досліджуваного об’єкта та середовища його функціонування, через що доцільно говорити лише про найзагальніші перспективи прогнозу.

Безпосередньо для розробника прогнозу вельми актуальне питання ступеня формалізації методу прогнозування. За цією ознакою всі методи прогнозування поділяються на дві категорії: евристичні та математичні, в основі кожної з яких лежить один з двох альтернативних підходів до розв’язку задач прогнозування.

Евристичні методи (від грецького слова "heurëka" – робити відкриття, або від більш відомої, завдяки історичному анекдотові, репліки Архімеда “Еврика!” – знайшов) започатковані ще в Древній Греції та Римі. Ці методи не потребують формального визначення проблеми і, відповідно, побудови алгоритму, що дає строгий й цілком інтерпретований розв’язок. Визначальна риса евристичного прогнозування – активне використання в тій чи іншій формах знань, досвіду та інтуїції експертів – висококваліфікованих фахівців з відповідної предметної галузі. Отриманий евристичний прогноз завжди інтегрований з конкретною особою – експертом чи групою експертів і має суб’єктивний характер.

В свою чергу евристичні методи поділяються на експертні (інтуїтивні) та аналітичні.

Експертні методи ґрунтуються виключно на оцінках експертів, зроблених відносно проблеми, яку вивчають. При цьому механізм продукування цих оцінок лишається невизначеним. Як правило, він невідомий навіть самому експерту, носить виключно індивідуальний, особистий характер і не може бути повторений чи відтворений кимсь іншим.

Натомість *аналітичні методи* прогнозування базуються на логічному (теоретичному та емпіричному) аналізі усього наявного масиву відомостей про об’єкт прогнозування, наслідком чого є побудова певної логічної схеми (моделі). Однак треба мати на увазі, що формування цієї логічної схеми звичайно проходить в умовах неповної визначеності щодо властивостей об’єкту прогнозування через завжди існуючий брак вхідної інформації. Недостатня інформація домислюється експертом відповідно до своїх вподобань та пріоритетів, тому отримана схема є суб’єктивною і, хоч дозволяє побудувати відповідну технологію прогнозування, придатну для багаторазового використання, отримані нею результати треба завжди сприймати як індивідуальний прогноз, зроблений виключно через призму особистості конкретного експерта.

Тому головна проблема евристичних методів – це визначення і врахування під час обробки експертних оцінок рівня компетентності експерта, його здатності робити достовірні припущення (прогнози) щодо

конкретних показників та перспектив розвитку об’єкту (процесу) прогнозування.

Математичне прогнозування базується на використанні математичної моделі об’єкту (процесу) прогнозування, побудованої за ретроспективними даними (тобто, отриманими до певного моменту часу в минулому), для обчислення його характеристик (станів) в довільний заданий момент.

Аргументами математичної моделі можуть бути контрольовані фактори (інформаційні ознаки), зміни кількісних значень яких впливають на значення вихідної змінної. Якщо до складу факторів входить час, об’єкт прогнозування належить до *класу динамічних об’єктів*, якщо тільки час – до часових рядів.

Сам процес математичного прогнозування можна умовно поділити на *чотири головних етапи*:

- збирання та підготовка вихідних даних;
- пошук та обґрунтування структури математичної моделі об’єкту (процесу) прогнозування (задача структурної ідентифікації моделі);
- оцінювання параметрів моделі за вихідними даними (задача параметричної ідентифікації моделі);
- перевірка якості моделі та прогнозування.

Ключовим етапом у наведеній процедурі є другий, зміст якого треба було б доповнити одним обов’язковим припущенням, невиконання якого лишає прогноз будь-якого сенсу: модель процесу (об’єкту) побудована за результатами аналізу ретроспективних даних, не змінюється впродовж часу упередження, тобто з поточного моменту до моменту, для якого обчислюють прогноз. Якщо це припущення виконується, точність прогнозу цілком залежить від рівня адекватності моделі об’єкту прогнозування, тобто від якості розв’язання задачі структурно-параметричної ідентифікації моделі.

Через те, що для моделі неминучі певні спрощення порівняно з оригіналом (об’єктом моделювання), точність прогнозу є принципово обмеженою, у зв’язку з чим постає питання про можливий (чи доцільний) інтервал упередження (прогнозу). Як вже зазначалося, у більшості прикладних застосувань розрізняють довго-, середньо- чи короткострокові прогнози. Для останнього звичайно (якщо це можливо в принципі) використовують математичне моделювання. Евристичні методи використовують для довгострокових прогнозів. Крім того, їх можна використовувати у разі якісної форми опису об’єкта прогнозування, тоді як математичні для побудови кращої моделі потребують кількісної форми подання вихідних даних.

В цьому випадку як вхідні дані, так і обраховані прогнозні (модельні, вихідні) значення мають кількісну форму представлення. Тому група

відповідних методів, що забезпечують отримання кількісного прогнозу, дістала назву *методів кількісного прогнозування*. Типовими представниками цієї групи є методи прогнозування часових рядів – впорядкованої у часі множини дискретних відліків певного параметру, величина якого змінюється у часі. Метою прогнозування часових рядів є визначення майбутньої поведінки контрольованого параметру на певному інтервалі упередження, розташованому справа від відмітки поточного моменту на часовій координаті. У наступних розділах буде розглянуто ряд методів прогнозування часового ряду, які можна вважати традиційними для цього класу задач.

Як зазначалося вище, традиційний підхід до кількісного математичного прогнозування базується на умові однаковості, незмінності математичної моделі (так звана "жорстка" модель [20–23]), як на інтервалі аналізу, тобто на ретроданих, так і на інтервалі упередження. Це означає, що фактично розглядається задача екстраполяції ретроспективної тенденції розвитку процесу, представленого сукупністю елементів часового ряду, на майбутнє, що принципово виключає можливість прогнозу змін, "стрибків", будь-яких інших кардинальних відступів від сформованої за ретроданими "жорсткої" моделі ряду. Тому для реалізації середньо- та довгострокових прогнозів, а також прогнозів, що потребують якісної форми опису та представлення прогнозної інформації, потрібне застосування нових методологічних принципів та підходів. Спроби їх пошуку та застосування достатньо детально описані в літературі [16,17].

6.2. Часові ряди. Структура, моделі.

Однією з ознак, за якою здійснюється класифікація моделей, є врахування у моделі фактору часу. За цією ознакою моделі поділяються на статичні та динамічні. Перші описують лише ті закономірності та співвідношення між параметрами ОД, існування яких не залежить від часу (не змінюється у часі). Динамічні моделі дозволяють показати процес функціонування та розвитку ОД, аналізувати тенденції й закономірності цього розвитку, дослідити вплив сукупності різних факторів на механізм цього розвитку у часі. Проте процедура ідентифікації динамічних моделей висуває специфічні вимоги щодо формату подання вихідних даних: вони мають бути представлені у формі часових рядів.

Часовий або динамічний ряд – упорядкована у часі послідовність дійсних чисел $z_1, z_2, ..., z_t, ...$, що характеризує поведінку ОД впродовж терміну його спостереження. Чисельні значення кожного окремого елементу часового ряду $\{z_t\}$, $t = 1,2,...$ називають його *рівнями*. Їх вимірюють у послідовні моменти часу (звичайно через рівні проміжки, так

званий регулярний ряд), через що порядок розташування елементів часового ряду дуже суттєвий.

Часовий ряд – особливий тип даних, який містить інформацію про плинність значень станів або характеристик ОД у часі. Ряд послідовних елементів може відображати деяку закономірність, тенденцію у функціонуванні чи розвитку ОД, наприклад, попередні значення ряду впливатимуть на зміну поточних чи майбутніх його значень. Таким чином, припускається наявність зв'язків або залежності між елементами одного й того ж ряду, яку, зокрема, можна зафіксувати у формі ММ часового ряду. Якщо вірогідне збереження цих закономірностей (зв'язків, залежностей) в майбутньому, реальною є можливість на основі побудованої за ретроданими (тобто даними, що описують поведінку часового ряду в минулому) моделі часового ряду передбачити його значення на певний період упередження.

Звичайно при дослідженні часових рядів в їх структурі намагаються окремо виділити детерміновану y_t та стохастичну e_t складові, процедури моделювання та прогнозування яких принципово відмінні. Найчастіше при цьому спираються на адитивну модель представлення часового ряду:

$$z_t = y_t + e_t \ , \quad t = \overline{1,n} \ . \tag{39}$$

де стохастична складова часто традиційно асоціюється із випадковою похибкою E , у відліках якої відсутня автокореляція:

$$\operatorname{cov}\{e_i, e_j\} = \begin{cases} 0, & i \neq j, \\ \sigma_e^2, & i = j, \end{cases}$$

а математичне очікування якої $M\{e\} = 0$.

В свою чергу в детермінованій складовій традиційно виділяють три адитивних компонента: тренд tr_t , сезонну s_t та циклічну c_t , тобто

$$y_t = tr_t + s_t + c_t \, . \tag{40}$$

Ці компоненти не мають єдиних загальноприйнятих визначень й змістовно залежать від конкретної предметної сфери застосування (іноді трендом називають загалом всю детерміновану складову моделі). В соціальних дисциплінах названі компоненти найчастіше ідентифікуються наступним чином.

Тренд – базовий (основний) детермінований компонент часового ряду, який характеризує вплив довгострокових постійно діючих факторів, повільний та інерційний.

Сезонний компонент відображає достатньо поширену як серед природних, так і штучних процесів регулярну повторюваність явищ: метеорологічних, кліматичних, соціально-економічних, демографічних й т.п., які не мають точної періодичності та сталості амплітуд. Приклади: зміни кожноденної завантаженості міського пасажирського транспорту,

коливання трафіку (кількість з'єднань) у телефонній мережі протягом доби, сезонна регулярність термінів проведення сільськогосподарських робіт ін.

Циклічний компонент описує нерегулярні, відносно довгі за часом періоди зростання та спаду рівнів часового ряду: зміни рівня радіації Чорнобиля, зміни чисельності популяції під час та після епідемії, зміни показників економіки внаслідок соціально-економічних спадів. Циклічний компонент важко ідентифікувати лише формальними засобами, необхідно залучати додаткову інформацію.

У аналізі часових рядів головну увагу звичайно приділяють виділенню трендового компонента, вважаючи його основою детермінованої складової часового ряду, на яку накладаються інші структурні компоненти моделей (39), (40). Тому далі в цьому розділі будуть розглядатися питання, пов'язані виключно з ідентифікацією трендової моделі часового ряду.

Залежно від сфери прикладного застосування накопичено певний досвід використання математичних моделей та визначено певну сукупність найбільш поширених базових трендових моделей. Зокрема, це показникові, степеневі, обернені та інші моделі [16]:

$$y = ab^t,\ \ y = at^b,\ \ y = ae^{bt},\ \ y = a + b\frac{1}{t},\quad y = \frac{1}{ab^t + c},$$

$$y = \frac{a}{1 + ae^{-bt}},\ \ y = \frac{c}{a + bt},\ldots .$$

Однак, якщо типової базової форми моделі заздалегідь визначити не вдається, найбільш гнучкою з точки зору її апроксимативних можливостей і наступного використання є *поліноміальна модель* виду:

$$y = a_0 + a_1 t + a_2 t^2 + \ldots + a_p t^p, \qquad (41)$$

де фактором, що впливає на залежну змінну y, є час $t = 1,2,\ldots,n$, який входить до моделі в різних степенях у вигляді відповідних регресорів: $t^1, t^2, \ldots, t^p$. В поліноміальній моделі звичайно використовується дискретний час *t*, значення якого співпадають з натуральним рядом чисел, тобто цей час слід вважати фіктивною змінною, що в моделях задається безвідносно до реального обліку часової змінної.

7. Прогноз трендового компонента часового ряду

7.1. Поліноміальні трендові моделі. Оцінювання якості, критерій Дарбіна-Уотсона.

Поліноміальна трендова модель (41) фактично є рівнянням

множинної регресії (11), тому для її ідентифікації цілком застосовні розглянуті в першій частині цього видання методи та процедури регресійного аналізу. Зокрема розширена матриця даних для моделі (41) матиме вигляд:

$$[Y+E,T]=[Z,T]=\begin{bmatrix} z_1 & 1 & 1 & 1 & \dots & 1 \\ z_2 & 1 & 2 & 4 & \dots & 2^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ z_t & 1 & t & t^2 & \dots & t^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ z_n & 1 & n & n^2 & \dots & n^p \end{bmatrix}=\begin{bmatrix} z_1 & \tau_1 \\ z_2 & \tau_2 \\ \dots & \dots \\ z_t & \tau_t \\ \dots & \dots \\ z_n & \tau_n \end{bmatrix}, \qquad (42)$$

а вектор МНК-оцінок параметрів визначатиметься співвідношенням:

$$\tilde{A}=[\tilde{a}_0,\tilde{a}_1,...,\tilde{a}_p]^T=(T^TT)^{-1}T^TZ. \qquad (43)$$

Проте на етапах вибору варіанту ММ та її верифікації виникає, завдяки специфічним особливостям часових рядів, можливість використання додаткових способів оцінювання якості варіантів ММ. Найчастіше це стосується аналізу нев'язок (залишків, похибок) $\varepsilon_t=z_t-\tilde{y}_t$, $t=\overline{1,n}$ трендової моделі, де $\tilde{y}_t=\tau_t\tilde{A}=[1,t,t^2,...,t^p]\tilde{A}$ – оцінка значення *t*-ого рівня часового ряду, розрахована за поліноміальною трендовою моделлю (41).

Зокрема, зважаючи на характер обмежень, що накладаються на властивості випадкової складової e_t рівня часового ряду, а саме, на припущення щодо некорельованості її значень:

$$\operatorname{cov}\{e_i,e_j\}=\begin{cases} 0, & i\neq j, \\ \sigma_e^2, & i=j, \end{cases}$$

видається доцільним перевірити, як виконується це обмеження для нев'язок $\varepsilon_1,\varepsilon_2,...,\varepsilon_t$, що за певних обставин сприймаються у якості оцінок відповідних значень випадкової складової *E*. Найбільш відомим і поширеним тестом перевірки моделі на наявність автокореляції між нев'язками є *критерій Дарбіна-Уотсона*. Критерій застосовується у тих випадках, коли тренд визначається для всього ряду [1,26,27].

Змістовну суть критерію можна пояснити наступним чином. У разі правильної специфікації трендової моделі за даними всього часового ряду, тобто коли структура функції тренду адекватно відображає динаміку змін значень рівнів часового ряду на інтервалі його спостереження, послідовність нев'язок $\varepsilon_1,\varepsilon_2,...,\varepsilon_t$ дійсно слід вважати оцінками значень випадкової складової *E*. Очевидно, що в цьому випадку автокорельованість нев'язок має бути мінімальною. У протилежному випадку, коли отримана трендова модель є недостатньо точною, в нев'язках присутня, крім

випадкової складової e_t, частка детермінованої складової y_t часового ряду [1], яка має високий рівень автокореляції. Тому, наприклад, серія позитивних або негативних знаків у послідовності нев'язок є своєрідним індикатором того, що побудована модель неадекватна вихідним даним. Більш вагомим статистичним підтвердженням цього факту є існування автокореляцій у серії нев'язок, одна з основних причин появи якої – наявність в них залишкового тренду.

Значення статистики Дарбіна-Уотсона розраховується за формулою:

$$d = \sum_{t=1}^{n-1} (\varepsilon_{t+1} - \varepsilon_t)^2 \Big/ \sum_{t=1}^{n} \varepsilon_t^2 . \qquad (44)$$

З аналізу формули (44) витікає, що якщо між значеннями $\varepsilon_1, \varepsilon_2, ..., \varepsilon_t$ присутня сильна позитивна автокореляція $r_1 = 1$, де r_1 – вибірковий коефіцієнт автокореляції між двома послідовними нев'язками часового ряду, то величина $d \approx 0$, якщо маємо сильну негативну автокореляцію $r_1 = -1$, то $d \approx 4$. При відсутності автокореляції $r_1 = 0$, статистика $d \approx 2$. Таким чином, значення d-статистики Дарбіна-Уотсона знаходяться у межах від 0 до 4 ($0 \le d \le 4$). На рис. 3 зображено схему розташування граничних значень d-статистики:

Рис.3. Зони розташування граничних значень *d*-статистики

Крім формули (44) для приблизного розрахунку статистики Дарбіна-Уотсона можна використати простішу формулу:

$$d \approx 2(1 - r_1) . \qquad (45)$$

Співвідношення (45) дає найбільш прозору інтерпретацію змісту критерію Дарбіна-Уотсона та його кількісним характеристикам.

Практичне використання критерію Дарбіна-Уотсона засновано на порівнянні величини d, розрахованої за формулою (44), з теоретичними значеннями d_1 та d_2, які показані на рис.3:

1)при $d_2 \le d \le (4 - d_2)$ – приймається гіпотеза про відсутність автокореляції між значеннями випадкової величини;

2)при $0 < d < d_1$ – виявляється позитивна автокореляція;

3)при $(4 - d_1) < d < 4$ – виявляється негативна автокореляція;

4)при $d_1 < d < d_2$ або $(4 - d_2) < d < (4 - d_1)$ – немає достатньої інформації для прийняття рішень, тобто отримуємо зону невизначеності.

Теоретичні граничні значення d_1 та d_2 для рівня значущості 5% і різних степенів поліному p представлено у табл. 2 (більш повні дані наведено у додатках в [3,4]).

Таблиця 2. Граничні значення d_1 та d_2 (p – степінь поліному, n – об'єм вибірки)

n	p=1		p=2		p=3		p=4	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74
40	1,44	1,54	1,39	1,6	1,34	1,66	1,29	1,72

Для визначення за відомою (вже ідентифікованою) поліноміальною моделлю виду (41) прогнозних значень тренду достатньо ввести до цієї моделі замість змінної t її кількісне значення l, в якому враховано інтервал упередження L. Так, якщо цей інтервал дорівнює одному, двом, трьом,... відлікам (крокам), тобто L=1,2,3,..., то значення змінної t кількісно дорівнюватиме $l = n + L$.

В матричній формі оцінка прогнозу тренду визначається виразом

$$\tilde{y}_l = \tau_l \tilde{A}, \qquad (46)$$

де $\tau_l = [1, l, l^2, ..., l^p]$ – вектор значень регресорів моделі (41) у точці прогнозу.

Помилка отриманої оцінки трендового прогнозу $\tilde{y}_l$ матиме (за умови адекватності структури моделі вихідним даним та обчисленню коефіцієнтів поліноміальної моделі методом найменших квадратів) дисперсію

$$\sigma^2\{\tilde{y}_l\} = \tau_l (T^T T)^{-1} \tau_l^T \sigma_e^2. \qquad (47)$$

Слід зазначити, що процедура виділення тренду суттєво ускладнюється в разі наявності в даних часового ряду циклічного компоненту (особливо в коротких чи середніх за обсягом рядах). Зокрема, помилкове включення рівнів, у складі яких присутній циклічний компонент, до сукупності даних, що використовується для ідентифікації трендової моделі, призводить до створення трендової залежності і робить неможливим її використання для прогнозу тренду часового ряду. В цьому випадку рекомендується [2] проаналізувати одночасну сукупну зміну у часі трендового та циклічного компонентів, відділивши від них шумову складову e_t, $t = \overline{1,n}$ шляхом згладжування даних часового ряду.

7.2. Згладжування часових рядів зваженим ковзким середнім. Критерій Аббе.

Мета згладжування часового ряду – виділення його детермінованої складової y_t, $t=\overline{1,n}$. Одним з найбільш поширених способів згладжування часового ряду є його обробка із застосуванням процедури зваженого ковзкого середнього. В загальному вигляді процедуру обчислення згладженої оцінки $\tilde{y}_t$ для *t*-ого рівня ряду z_t описує вираз:

$$\tilde{y}_t = \sum_{j=-m}^{m} v_j z_{t+j}, \qquad (48)$$

де v_j, $j=\overline{-m,m}$ – вагові коефіцієнти згладжуючого фільтру $Z_{KI}=[z_{t-m},...,z_t,...,z_{t+m}]^T$ – рівні ряду, що утворюють так званий ковзкий інтервал (КІ) довжиною 2*m*+1, який отримав назву завдяки своєму покроковому переміщенню вздовж часового ряду у ході виконання згладжування останнього. Покрокове переміщення КІ здійснюється шляхом відкидання його крайнього лівого елементу та приписування до КІ справа найближчого сусіднього рівня часового ряду. При такому зміщенні КІ черговий згладжуваний рівень ряду опиняється у середині КІ, тобто оцінка $\tilde{y}_t$ завжди обчислюється для середини КІ. Через це *m* перших та *m* останніх рівнів часового ряду залишається незгладженими. Це явище отримала назву крайового ефекту.

Для отримання згладженої оцінки за даними КІ будується поліноміальна модель виду (41), яка має апроксимувати фрагмент Z_{KI} часового ряду в межах ковзкого інтервалу. Для цієї моделі розраховується вектор МНК-оцінок параметрів $\tilde{A}_{KI}$, який далі використовується для обрахування згладженої оцінки в середній точці КІ. Отримана оцінка є шуканою згладженою оцінкою *t*-ого рівня часового ряду. За аналогією з (46) маємо:

$$\tilde{y}_t = \tau_{L\,KI}\tilde{A}_{KI} = \tau_{L\,KI}(T_{KI}^T T_{KI})^{-1}T_{KI}^T Z_{KI}, \qquad (49)$$

де $\tau_{L\,KI}$ – середній рядок матриці плану T_{KI}.

Зважаючи, що довжина 2*m*+1 ковзкого інтервалу та порядок *p* поліному (41) залишаються незмінними впродовж всієї обробки часового ряду, незмінною буде і матриця плану T_{KI}, тому матричний добуток у правій частині виразу (49), пов'язаний лише з параметрами *m* та *p*, можна обрахувати заздалегідь:

$$V_{KI} = [v_{-m},...,v_m] = \tau_{L\,KI}\,(T_{KI}^T T_{KI})^{-1} T_{KI}^T . \qquad (50)$$

Зробивши відповідну заміну у співвідношенні (49):

$$\tilde{y}_t = V_{KI}\, Z_{KI},$$

отримуємо матричну форму запису виразу (48), тобто V_{KI} – вектор вагових коефіцієнтів згладжуючого фільтру. Властивість сталості значень вагових коефіцієнтів при заданих параметрах m та p фільтрів дозволяє виконати попереднє обрахування ваг для різних сполучень m та p та внести отримані значення векторів $V(m, p)$ до довідкових розділів спеціалізованої літератури [4].

Для практичної реалізації процедури згладжування актуальне питання вибору параметрів m, p згладжуючого фільтру. Звичайно воно розв'язується шляхом підбору значень m, p за допомогою критерію якості згладжування. Методичною базою цього перебору може бути викладена у розділі 7.1 ідея перевірки автокорельованості нев'язок $\varepsilon_t = z_t - \tilde{y}_t$, $t = \overline{m+1, n-m}$. Саме ця ідея реалізована у *критерії Аббе*, який можна застосувати для вибору параметрів фільтру:

$$AB = \sum_{t=1}^{n-1} (\varepsilon_{t+1} - \varepsilon_t)^2 \Big/ 2\sum_{t=1}^{n} \varepsilon_t - \bar{\varepsilon})^2 , \qquad (51)$$

де $\bar{\varepsilon} = \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t$ – середнє арифметичне значення нев'язок.

Якщо $AB > AB_{meop}(n)$, то приймається гіпотеза некорельованості нев'язок, тобто якісного згладжування. Теоретичне значення $AB_{meop}(n)$ для n>60 розраховується за формулою:

$$AB_{meop}(n) = 1 + u_\alpha / \sqrt{n + 0{,}5(1 + u_\alpha^2)}, \qquad (52)$$

де u_α – α-квантиль нормованого нормального розподілу. При $n \le 60$ значення $AB_{meop}(n)$ для трьох найбільш застосовних рівнів значущості α наведено в [18].

Зазначимо, що оцінку дисперсії нев'язки, отриману для найближчого до 1 значення статистики Аббе, можна вважати найбільш точною оцінкою дисперсії σ_e^2 стохастичної складової e_t часового ряду:

$$\sigma_e^2 \approx \sum_{t=1}^{n} (\varepsilon_t - \bar{\varepsilon})^2 / (n-1). \qquad (53)$$

7.3. Оцінка точності прогнозу за ретроданими

Нажаль розглянуті вище критерії Дарбіна-Уотсона (44) та Аббе (51) дають достатньо чітке уявлення про апроксимативні якості моделей на інтервалі спостереження, але не відповідають на питання про якість прогнозу. Більш-менш об'єктивну відповідь на це питання можна отримати використовуючи ретропрогноз, тобто прогноз, знайдений за трендовою моделлю, побудованою для фрагменту часового ряду $\{z_t\}$, $t=\overline{1,n_1}$, $n_1 = n - L - n_2$, де L – інтервал прогнозу (звичайно $L = 1 \div 3$), $n_2 \le 0{,}5n$, при цьому було б бажано, щоб для n_1 та n_2 за можливістю виконувались умови: $n_1 \ge 0{,}5n$, n_2>10. Розрахованому в цьому випадку прогнозному значенню рівня $\tilde{y}_{n_1+L}$ слід співставити відомий елемент z_{n_1+L} вихідного ряду, визначивши таким чином нев'язку ретропрогнозу, що складатиме $\varepsilon_{R,n_1+L} = z_{n_1+L} - \tilde{y}_{n_1+L}$, де нижній індекс R означає приналежність до ретропрогнозу. Наступна прогнозна оцінка $\tilde{y}_{n_1+L+1}$ розраховується за подовженим на одиницю фрагментом ряду $\{z_t\}$, $t=\overline{1,n_1+1}$, відповідна нев'язка дорівнюватиме $\varepsilon_{R,n_1+L+1} = z_{n_1+L+1} - \tilde{y}_{n_1+L+1}$. Подібні підрахунки продовжуватимуться доти, доки на фрагменті $\{z_t\}$, $t=\overline{1,n-L}$ не буде обчислений останній ретропрогноз $\tilde{y}_n$. Слід наголосити, що при ідентифікації моделей на фрагменті часового ряду структура цих моделей має співпадати із структурою трендової моделі, визначеною за даними всього часового ряду (відповідно до розділу 7.1).

За отриманим рядом нев'язок ретропрогнозу ε_{R,n_1+L}, ε_{R,n_1+L+1},..., $\varepsilon_{R,n}$ можна проаналізувати їх сталість або наявні тенденції до змін, спробувати оцінити дисперсію прогнозної оцінки $\tilde{y}_{n+L}$. Для цього знайдемо оцінку дисперсії нев'язок ретропрогнозу. В якості її можна прийняти середній квадрат рівнів наведеного вище ряду нев'язок або оцінку, обраховану для цього ряду методом експоненціального згладжування. В останньому випадку, застосовуючи рекурентну процедуру

$$\sigma^2_{R,\varepsilon,t} = \alpha\varepsilon^2_{R,t} + (1-\alpha)\sigma^2_{R,\varepsilon,t-1}, \quad t=\overline{m+L,n}, \qquad (54)$$

де в якості початкової оцінки $\sigma^2_{R,\varepsilon,t-1}$ приймається значення σ^2_{ε}, розраховане за формулою (47), отримуємо оцінку дисперсії ретропрогнозу $\sigma^2_{R,\varepsilon,n}$ для кінця часового ряду.

Зважаючи на те, що дисперсія ретропрогнозу складається з двох частин, одна з яких – дисперсія стохастичної складової часового ряду σ^2_e, а

друга – саме дисперсія прогнозу $\sigma^2_{\varepsilon,L}$ на інтервалі упередження L, знаходимо:

$$\sigma^2_{\varepsilon,L} = \sigma^2_{R,\varepsilon,n} - \sigma^2_e, \qquad (55)$$

де σ^2_e визначається за формулою (53).

7.4. Метод гармонічних вагових коефіцієнтів

Метод гармонічних вагових коефіцієнтів був запропонований польським статистиком З.Хелвигом для прогнозування часових рядів, що не мають сезонних та циклічних коливань. Основна ідея методу – обробка рівнів часового ряду таким чином, що більш пізнім його рівням надається більша вага. Це досягається через введення у процедуру обробки спеціальних вагових коефіцієнтів.

Переваги методу гармонічних ваг у порівнянні з іншими методами, де також використовуються зважування рівнів часового ряду, у тому, що його застосування не потребує ніяких припущень щодо виду тренду. Розглянемо цей метод більш докладно.

Нехай маємо часовий ряд z_t, $t = 1,2,...,n$, який можна розкласти на невипадкову функцію від часу (тренд) та стаціонарний випадковий компонент: $z_t = tr_t + e_t$. Треба визначити оцінку прогнозу рівня $\tilde{y}_t$ для $(n+L)$-ого елемента часового ряду, де $L = 1,2,3$ – інтервал упередження. Процедура прогнозу рівня $\tilde{y}_t$ складається з трьох етапів: згладжування вихідного часового ряду $\{z_t\}$, обчислення середньозваженого приросту згладжених значень $\tilde{y}_t$, $t = \overline{1,n}$ на інтервалі, який дорівнює інтервалу упередження $L = 1,2,3$, визначення оцінки прогнозу.

Згладжування вихідного часового ряду $\{z_t\}$, зазвичай реалізується методом зваженого ковзкого середнього першого порядку (p=1) з ковзким інтервалом (КІ) довжиною три або п’ять елементів (тобто m=1 або m=2). Для запобігання скороченню згладженого часового ряду на $2m$ елементів внаслідок так званого крайового ефекту (відкидання m перших та m останніх елементів ряду при застосуванні для обчислення згладжених оцінок формули 48)) використаємо формулу

$$\tilde{y}_{t+k} = \sum_{j=-m}^{m} v_{j,k} z_{t+j}, \qquad (56)$$

в якій на відміну від формули (48) використовуються вагові коефіцієнти $v_{j,k}$, $j = \overline{-m,m}$ з подвійною індексацією, що дозволяє, залежно від значення k $(k = \overline{-m,m})$, обчислити згладжені оцінки для будь-якого

елемента КІ. Використовуючи цю можливість, за формулою (56) для m=1 при $t=2,\ k=-1$ обраховується згладжена оцінка першого рівня часового ряду, при $t=n-1$, k =1 – останнього, у випадку m=2 при $t=3,\ k=-2,-1$ – дві перші оцінки та при $t=n-2,\ \ k=1,2$ – відповідно дві останні. При k=0 обчислюються згладжені оцінки у середині КІ, тобто формула (56) співпадає із (48).

Обчислимо послідовність приростів для пар рівнів, розділених L періодами:

$$\omega_{t+L}=\tilde{y}_{t+L}-\tilde{y}_t,\quad t=\overline{1,n-L}, \tag{57}$$

У відповідність кожному приросту поставимо гармонічну вагу, яка обчислюється за формулою:

$$m_{t+L}=\sum_{j=L}^{n-1}\frac{1}{n-j},\quad t=\overline{1,n-L} \tag{58}$$

або через рекурентну процедуру, що задається співвідношенням:

$$m_{t+L+1}=m_{t+L}+\frac{1}{n-t},\ \text{ де }\ m_{1+L}=\frac{1}{n-L}. \tag{59}$$

Сума гармонічних ваг дорівнюватиме

$$S(L)=\sum_{t=1}^{n-L}m_{t+L}=n-L. \tag{60}$$

Вводячи нормуючий множник $1/S(L)$, розрахуємо сукупність гармонічних коефіцієнтів c_{t+L}^{L}:

$$c_{t+L}^{L}=\frac{m_{t+L}}{n-L},\quad t=\overline{1,n-L}, \tag{61}$$

сума яких дорівнюватиме одиниці (так звана умова незміщеності системи коефіцієнтів):

$$\sum_{t=1}^{n-L}c_{t+L}^{L}=1; \tag{62}$$

Розрахуємо середньозважені прирости для різних значень L:

$$\overline{\omega}_{t+L}=\sum_{t=1}^{n-L}c_{t+L}^{L}\omega_{t+L}, \tag{63}$$

Прогнозні значення рівнів ряду з інтервалом упередження L обчислюються за формулою:

$$\tilde{y}_{n+L}=\tilde{y}_n+\overline{\omega}_{t+L}, \tag{64}$$

Можливі варіанти виразів для обчислення рівнів прогнозів на інтервалах L=1,2,3 наведено нижче у таблиці.

1-й варіант	2-й варіант	3-й варіант

$\tilde{y}_{n+1} = \tilde{y}_n + \bar{\omega}_{t+1}$		
$\tilde{y}_{n+2} = \tilde{y}_{n+1} + \bar{\omega}_{t+1}$	$\tilde{y}_{n+2} = \tilde{y}_n + \bar{\omega}_{t+2}$	
$\tilde{y}_{n+3} = \tilde{y}_{n+2} + \bar{\omega}_{t+1}$	$\tilde{y}_{n+3} = \tilde{y}_{n+1} + \bar{\omega}_{t+2}$	$\tilde{y}_{n+3} = \tilde{y}_n + \bar{\omega}_{t+3}$

8. Прогнозування стохастичної складової часового ряду. Метод авторегресійних моделей.

8.1. Модель авторегресії: основні поняття

При наявності лінійної залежності між випадковими величинами в послідовності $E_1, E_2, ..., E_t, ...$ зустрічаємося з явищем *авторегресії* – регресії, пояснюючими змінними якої є попередні рівні динамічного ряду [1,2,3].

Моделлю авторегресії називається модель стаціонарної послідовності, що відображає значення показника y_t у вигляді лінійної комбінації скінченого числа попередніх значень цього показника та адитивної випадкової складової:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + e_t. \qquad (65)$$

*Лаг (*або *часовий лаг, лаг запізнення)* – проміжок часу τ, за який зміна аргументу призведе до зміни результативного показника. Наявність запізнення означає, що вплив змінної *x* на змінну *y* не проявляється негайно, а розтягується на деякий проміжок часу. Значення *p* в моделі (65) визначає порядок авторегресії. Якщо в моделі часового ряду використовується дискретний час *t*, значення якого співпадають з натуральним рядом чисел, максимальний лаг визначає порядок авторегресії *p*.

Розглянемо, чим може бути корисна модель авторегресії для прогнозу часових рядів. Як зазначалося у п. 6.2, часовий ряд z_i, $i = 1,2,...,n$ можна розкласти на детерміновану функцію часу y_t, котра частіше за все асоціюється із *трендом*, та випадкову складову e_i.

Припустимо, що, приймаючи введені в п. 6.2 певні обмеження щодо властивостей складової e_i в елементах часового ряду $\{z_i\}$, $i = \overline{1,n}$, її можна вважати реалізацією дискретного білого шуму, отриманого внаслідок дискретизації "чисто випадкового" стаціонарного стохастичного процесу з нульовим математичним очікуванням. Тоді $z_i = y_i + e_i$, де y_i – інформативна складова *i*-ого елемента часового ряду. Обчислені після виділення тренду нев'язки ε_i, $i = \overline{1,n}$ в загальному випадку відрізняються

від e_i, бо в своєму складі містять ще один додатковий компонент $\delta_i = y_i - \tilde{y}_i$ – помилку апроксимації трендовою моделлю інформативної складової y_i часового ряду. Частіше за все δ_i можна інтерпретувати як наслідки присутності в часовому ряді певної сукупності мікроциклових компонентів, що не можуть бути описані трендовою моделлю. Тому фрагменти сукупності $\{\delta_i\}$, $i = \overline{1,n}$ являють собою сильнокорельовані послідовності, які за своїми частотно-часовими властивостями значно ближче до рядів $\{y_i\}$, $\{\tilde{y}_i\}$, ніж до нев'язки $\{\varepsilon_i\}$, з елементами якої їх зв'язує співвідношення

$$\varepsilon_i = \delta_i + e_i, \quad i = \overline{1,n}, \tag{66}$$

тобто значення нев'язки за певних обставин можна вважати оцінками компоненту δ_i .Зважаючи на це, якщо припустити, що послідовність $\{\delta_i\}$ апроксимується моделлю виду

$$\delta_i = a_1\delta_{i-1} + a_2\delta_{i-2} + ... + a_p\delta_{i-p}, \tag{67}$$

і це припущення виявиться слушним, то, підставивши до правої частини моделі відповідні значення $\varepsilon_{i-1}, \varepsilon_{i-2}, ..., \varepsilon_{i-p}$ (які у даному випадку виступають у якості оцінок рівнів ряду $\{\delta_i\}$), можна обрахувати оцінки корегуючої поправки $\tilde{\delta}_i$, додавання якої до трендового прогнозу $\tilde{y}_i$ дозволить підвищити точність останнього. Таким чином загальний прогноз часового ряду буде складатися з результатів двох прогнозів:

$$\tilde{\tilde{y}}_i = \tilde{y}_i + \tilde{\delta}_i. \tag{68}$$

Тобто за допомогою авторегресійної моделі (65) з нев'язок трендової моделі $\{\varepsilon_i\}$, $i = \overline{1,n}$ можна виділити додаткову корисну інформацію щодо апріорно невідомої інформативної складової y_i елементів часового ряду. Однак для цього слід попередньо визначити структуру моделі авторегресії (65), зокрема її порядок p. Відомості про це можна отримати з аналізу автокореляційної функції нев'язки $r(k)$, точніше з вибіркових оцінок значень цієї функції r_k, розрахованих відповідно для дискретних моментів часу k=0,1,2,... .

Для обчислення корегуючої поправки оцінок тренду часового ряду за допомогою авторегресійних моделей необхідно виконати ряд послідовних операцій:

- виключити тренд $\tilde{y}$, уточнивши структуру моделі тренду відомими способами;
- розрахувати відхилення (нев'язки) ε_i моделі тренду;
- для часового ряду $\{\varepsilon_i\}$ розрахувати автокореляційні функції r_k;
- знайти порядок p авторегресійної моделі та оцінити її параметри

$a_1, a_2, \ldots$;

– використовуючи отримані оцінки параметрів авторегресійної моделі, побудувати корегуючу поправку трендового прогнозу часового ряду;
– оцінити ефективність введення корегуючої поправки.

Ці операції можуть повторюватися в процесі уточнення досліджуваних моделей, забезпечуючи за рахунок аналізу значень різноманітних показників властивостей даних та моделей логічні обґрунтування отриманих числових оцінок прогнозу при збереженні припущення про незмінність умов та характеристик часового ряду на період прогнозування.

8.2. Оцінювання структури та коефіцієнтів рівняння авторегресії

Процес авторегресії характеризується тим, що його автокореляційна функція $r(k)$ є затухаючою (рис. 4), на відміну від автокореляційної функції білого шуму, яка теоретично має дорівнювати 0 для всіх $k \neq 0$. При $k = 0$ $r(k)$ завжди дорівнює 1.

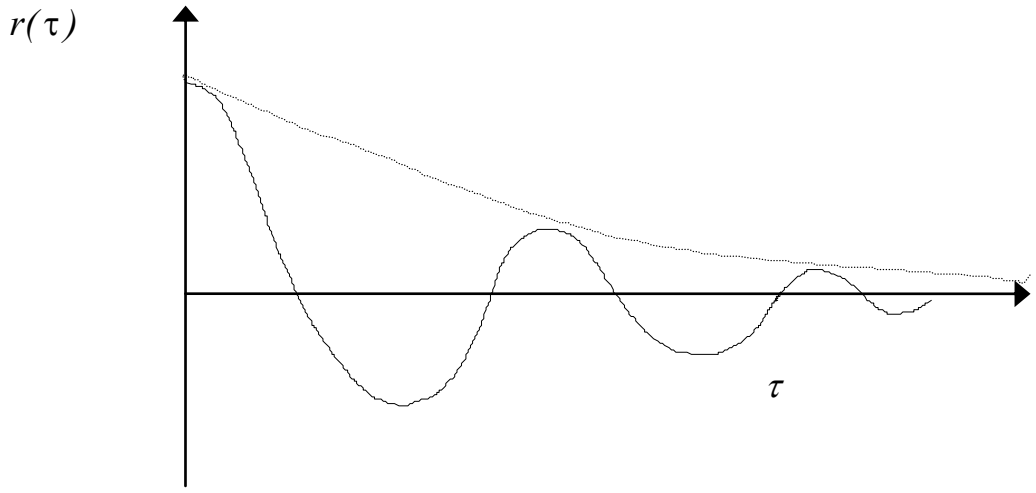

Рис.4. Приклади автокореляційних функцій стаціонарних випадкових процесів

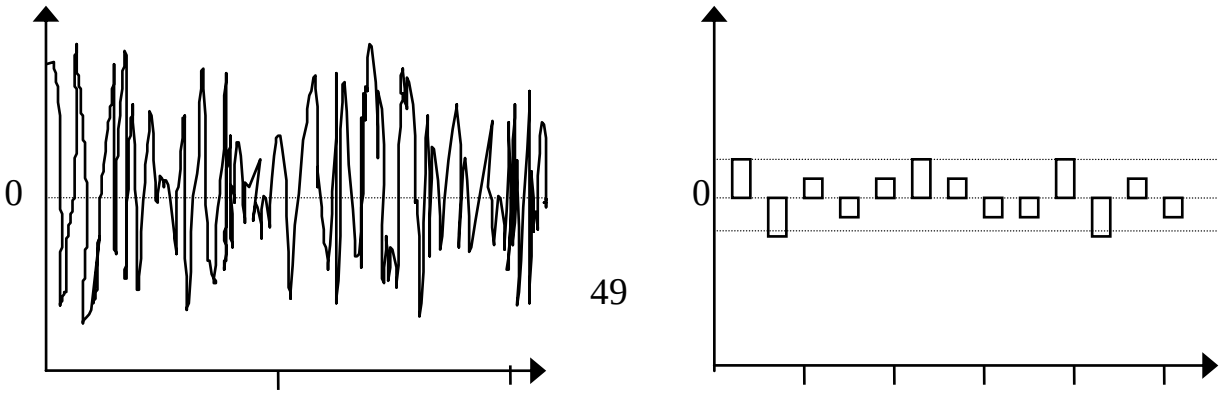

Рис. 5. Фрагмент реалізації низькокорельованого випадкового процесу (а) та оцінка його автокореляційної функції (б)

Графік вибіркових оцінок автокореляційної функції називається *корелограмою*, яка характеризує рівень лінійного зв'язку між елементами часового ряду, віддаленими один від одного відповідно на 0,1,2,... періоди опитування (дискретизації). Для побудови корелограми необхідно:

1) оцінити дисперсію нев'язки:

$$\sigma_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 ; \qquad (69)$$

2) розрахувати вибіркові значення кореляційної функції

$$r_k = \frac{\sum_{j=1}^{n-k}\varepsilon_j\varepsilon_{j+k}}{(n-k)\sigma_\varepsilon^2}, \quad k=\overline{1,n_1}, \qquad (70)$$

де n_1 задається рівним від третини до половини обсягу сукупності n.

На рис. 5а відображено вихідний ряд білого шуму та на рис. 5б – корелограму шуму. Для гаусовського білого шуму наближено [1,2] можна вказати 95% довірчий інтервал вибіркових оцінок значень r_k: $-1/n \pm 2/\sqrt{n}$. Його зображено на графіку корелограми (рис. 5б) двома тонкими горизонтальними лініями. Якщо вибіркові оцінки кореляційної функції не виходять за межі вказаного довірчого інтервалу, можна припустити, що дані, за якими розраховані оцінки r_k, є реалізацією дискретного білого шуму.

Одним із важливих етапів побудови авторегресивної моделі є визначення її порядку p та обчислення її параметрів. Якщо порядок моделі вибрано не вдало, то процедуру обчислення оцінок можна повторити для моделі авторегресії іншого порядку чи структури. Необґрунтоване підвищення порядку моделі та ускладнення її структури знижує точність оцінок параметрів та якість прогнозу. Водночас недостатня кількість коефіцієнтів моделі й занадто малий порядок авторегресії не дадуть можливість адекватно оцінити динаміку процесу та спрогнозувати його подальші зміни. Тому задача дослідника – вибрати модель авторегресії найменшого порядку за умови забезпечення достатньої точності опису

даних та прогнозування.

Для визначення порядку авторегресії використовують автокореляційну функцію r_k. Визначають значущі значення вибіркових оцінок r_i автокореляційної функції, старший індекс i яких приймається за порядок p авторегресійної моделі.

При цьому враховують, що вибіркові оцінки автокореляційної функції наближено [1] характеризуються зміщенням $-1/n$ і дисперсією $\sigma_r^2 = 1/n$ (середньоквадратичне відхилення $\sigma_r = 1/\sqrt{n}$). Якщо значення r_k не потрапляють до довірчого інтервалу $-1/n \pm 2\sigma_r$, тобто $-1/n \pm 2/\sqrt{n}$, вони вважаються значущими. У протилежному випадку, тобто коли

$$-1/n - 2/\sqrt{n} \le r_k \le -1/n + 2/\sqrt{n}, \qquad (71)$$

приймаємо гіпотезу $r(k) = 0$.

На рис. 4.3 штриховими лініями позначено область значущих значень оцінок r_k:

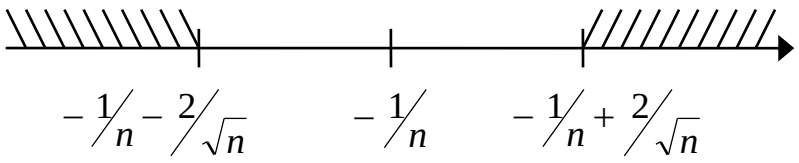

Рис.6. Інтервали значущих оцінок автокореляційної функції

Після орієнтовного визначення структури регресійної моделі для її подальшого уточнення необхідно виконати оцінювання параметрів авторегресії. Зокрема для моделі виду (65) це може бути реалізоване методом найменших квадратів (МНК). Приймаючи до уваги структуру моделі (65), сформуємо із масиву нев'язок $\{\varepsilon_i\}$ вектор зашумлених значень залежної змінної Z та матрицю плану X:

$$Z = \begin{bmatrix} \varepsilon_{p+1} \\ \varepsilon_{p+2} \\ \dots \\ \varepsilon_n \end{bmatrix}; \quad X = \begin{bmatrix} \varepsilon_p & \varepsilon_{p-1} & \dots & \varepsilon_1 \\ \varepsilon_{p+1} & \varepsilon_{p+2} & \dots & \varepsilon_2 \\ \dots & \dots & \dots & \dots \\ \varepsilon_{n-1} & \varepsilon_{n-2} & \dots & \varepsilon_{n-p} \end{bmatrix}. \qquad (72)$$

Обробка цих даних за МНК дає вектор МНК-оцінок параметрів $\tilde{A}$ авторегресійної моделі.

Однак як вже зазначалося, введена орієнтовна оцінка p порядку

авторегресії може значно перевищувати фактичне значення порядку, а структура моделі – містити надлишкові регресори. Тому необхідно обов'язково провести перевірку та уточнення структури моделі шляхом виявлення за критерієм Стьюдента незначущих параметрів та виключення із структури відповідних їм регресорів (див. п.5.3).

Кількість регресорів, що звичайно має бути вилучена з вихідної структури, може значно перевищувати число регресорів, які залишилися. Порядок вилучення припускає високий ступінь суб'єктивізму при прийнятті проміжних рішень. Через це результатом спрощення може бути кілька достатньо суперечливих кінцевих варіантів структури. Тому для визначення структури авторегресії рекомендується використовувати суттєво різні підходи [1,2,12]. Для отримання кращої структури моделі результати, отримані за різними підходами, треба співставити та проаналізувати. Викладений вище метод виключення регресорів із початкової структури моделі доцільно комбінувати з методами включення регресорів. Це вже розглянутий у п.3 метод крокової регресії, а також евристичний підхід, в основі якого включення до початкової структури моделі лагових змінних, лаг яких співпадає з індексами вибіркових коефіцієнтів $r_q, r_s, r_v, ...$, значення яких виходять за межі обрахованого вище 95% довірчого інтервалу для оцінок функції автокореляції (при чому для індексів виділених вибіркових коефіцієнтів має виконуватися умова: $1 \le q < s < v < ...$). Методи включення дозволяють ефективно контролювати процес ускладнення початкової моделі за допомогою критерію Фішера та коефіцієнта детермінації R^2, комбінуючи цей контроль із перевіркою значимості коефіцієнтів моделі за Стьюдентом. Цей процес триває доки не буде отримана модель, адекватна вихідним даним, яка і буде використана для коригування оцінок тренду. В загальному випадку, деталізуючи наведене вище співвідношення (66), що описує спосіб введення корекції тренду, з урахуванням моделі (65) отримуємо:

$$\tilde{\tilde{y}}_i = \tilde{y}_i + a_i \varepsilon_{i-1} + a_2 \varepsilon_{i-2} + ... a_p \varepsilon_{i-p} \qquad (73)$$

Об'єктивним показником ефективності використання побудованої моделі авторегресії має бути некорельованість значень нев'язки ε'_i, розрахованої за скорегованими оцінками тренду:

$$\varepsilon'_i = z_i - \tilde{\tilde{y}}_i, \quad i = \overline{p+1, n}. \qquad (74)$$

Тобто вибіркові коефіцієнти автокореляційної функції r_1, r_2, ..., розраховані за рядом $\{\varepsilon'_i\}$, мають перебувати в межах введеного вище 95% довірчого інтервалу.

9. Прогноз детермінованої складової часового ряду. Метод експоненціального згладжування.

9.1. Локальні моделі часового ряду

Відповідно до змісту адитивної моделі часового ряду (40), детермінована складова y_t цієї моделі в загальному випадку відрізняється від тренду tr_t, включаючи в себе інші компоненти, наявність яких обумовлює необхідність застосування для опису складової y_t певних локальних моделей. Особливість цих моделей полягає в тому, що в ряді потреби вони можуть змінювати свою структуру та параметри, адаптуючись до появи у складі y_t додаткових компонентів (зокрема циклічного компоненту c_t).

Припустимо, що інформативна складова y_t, $t=\overline{1,n}$ моделі (40) часового ряду отримана в результаті регулярної дискретизації процесу $y(\tau)$, який в околі точки t може бути розвинений у ряд Тейлора

$$y(\tau) = a_{0t} + a_{1t}(\tau - t) + \frac{1}{2!}a_{2t}(\tau - t)^2 + ... + \frac{1}{k!}a_{2t}(\tau - t)^k + ..., \qquad (75)$$

причому для цілком достатньої точності наближення можна обмежитися першими k+1 членами ряду, тобто поліном

$$P_t(\tau, k) = \sum_{i=0}^{k} \frac{a_{it}}{i!}(\tau - t)^i . \qquad (76)$$

Цей поліном $P_t(\tau, k)$, коефіцієнти якого визначені із врахуванням особливостей поведінки функції $y(\tau)$ в околі точки *t*, являє собою локальну модель часового ряду. Точкою локалізації цієї моделі є рівень z_t (або *t* –й момент часу). В принципі точкою локалізації може бути будь-яке значення часового ряду $\{z_t\}$, $t=\overline{1,n}$, що певним чином наближає локальну модель до моделі зваженого ковзкого середнього. Однак модель ковзкого середнього будується лише за даними, що входять до складу КІ, тоді як локальна модель – за даними фрагменту часового ряду $\{z_1, z_2, ..., z_t\}$, причому якщо $t = n$, то цей фрагмент – весь часовий ряд.

Одним з основних питань побудови локальної моделі є визначення вектора її коефіцієнтів $A = [a_{0t}, ..., a_{kt}]^T$. Ця задача розв'язується шляхом застосування зваженого МНК. Зокрема, у випадку $t = n$ коефіцієнти поліному $P_n(\tau, k)$ визначаються з умови мінімуму зваженої суми квадратів нев'язок виду:

$$\alpha \sum_{j=0}^{n-1} \beta^j [z_{n-j} - \sum_{i=0}^{k} \frac{a_{it}}{i!}(-j)^i]^2 = \alpha \sum_{j=0}^{n-1} \beta^j \varepsilon_{n-j}^2 = \min, \qquad (77)$$

де $\beta < 1$, $\alpha = 1 - \beta$. В результаті отримуємо локальну модель часового ряду, яка добре адаптується до його особливостей в околі точки $t = n$, виказуючи високі апроксимативні якості в цьому околі, однак поступово втрачає точність апроксимації з віддаленням від $t = n$. Цими властивостями модель завдячує вектору своїх коефіцієнтів $A_n = [a_{0n}, ..., a_{kn}]^T$, визначеному за критерієм (77), точніше системі ваг $\{w_j\}$, $j = \overline{1,n}$:

$$w_n = \alpha > w_{n-1} = \alpha\beta > w_{n-2} = \alpha\beta^2 > ... > w_{n-j} = \alpha\beta^j > ... \\ ... > w_1 = \alpha\beta^{n-1}, \qquad (78)$$

значення яких спадають за експоненціальним законом так, що квадрати нев'язок в околі $t = n$ мають найвищу вагу (чим й забезпечується найкраща апроксимація даних цієї області ряду), яка потім монотонно зменшується, наближаючись до нуля для початкових індексів елементів часового ряду.

Якщо до локальної моделі (76), побудованої для $t = n$, ввести значення змінної $\tau = t + L$, отримаємо прогноз часового ряду на інтервал упередження L. Вибір значення L залежить від ширини околу, в якому локальна модель $P_n(\tau, k)$ задовільно апроксимує часовий ряд. Окіл ширшає із збільшенням порядку k та зменшенням значення β і навпаки, вужчає із зростанням β та зменшенням k. Звичайно $k \le 2$, а $0{,}7 \le \beta < 1$, однак іноді ліва межа для параметру β може спадати навіть до 0,3.

У разі збільшення обсягу часового ряду (звичайно n зростає в результаті появи нових даних, що поповнюють часовий ряд справа), локальна модель $P_n(\tau, k)$ має постійно адаптуватися до нових даних, кожного ряду змінюючи значення вектора коефіцієнтів A_n відповідно до постійно обновлюваної статистики в лівій частині критерію (77).

Загалом наведена вище процедура застосування локальної моделі для прогнозу часового ряду видається достатньо копіткою, особливо в разі покрокового оновлення ряду за рахунок поелементного приєднання до нього справа нових даних. В цьому випадку виникає необхідність в багаторазовому застосуванні зваженого МНК для перерахунків значень вектора A_n на покроково зростаючій вибірці даних. Тому видається доцільним застосування рекурентних методів обробки часового ряду, які дозволяютьє суттєво спростити обчислення коефіцієнтів локальних моделей $P_t(\tau, k)$, $t = \overline{1,n}$. Найбільш поширеним серед них можна вважати метод експоненціального згладжування – модифікацію дисконтованого

методу зважених найменших квадратів, пристосовану для рекурентного прогнозу елементів часового ряду.

9.2. Експоненціальне згладжування

Ідея методу експоненціального згладжування [1,3] полягає в згладжуванні часового ряду ковзкою середньою з експоненціальними вагами (78). Така середня більше характеризує значення процесу на кінці інтервалу згладжування, ніж на початку. Справді, здебільшого на прогнозні значення суттєво впливають останні рівні часового ряду, тому їм надається більша вага порівняно з початковими, які, однак, не виключають зовсім з аналізу, тому що вони також несуть певну інформацію про досліджуваний процес. Назва методу випливає з того, що дані згладжуються за допомогою зваженої середньої, у якій ваги змінюються згідно з експоненціальним законом.

Ряд спостережень можна згладжувати будь-яку кількість разів, наприклад, ряд, згладжений один раз, можна згладити ще один раз ковзкою середньою з експоненціальними вагами (так зване подвійне експоненціальне згладжування), отриманий ряд знову згладжуємо ковзкою середньою з експоненціальними вагами (потрійне експоненціальне згладжування) тощо.

Метою багаторазового експоненціального згладжування є рекурентний розрахунок оцінок коефіцієнтів рівняння виду (75), який базується на обчисленні експоненціальних середніх (ЕС) різних порядків.

ЕС першого порядку розраховують за вихідною вибіркою $\{z_1, z_2, ..., z_n\}$:

$$S_t^{(1)} = \alpha z_t + \beta S_{t-1}^{(1)}, \qquad (79)$$

де α - постійна згладжування (звичайно α обирається у діапазоні $0{,}03 \div 0{,}4$, але іноді може зростати до $0{,}7 \div 0{,}8$), $\beta = 1 - \alpha$ – фактор затухання.

Використання рекурентної формули (79) потребує завдання початкового значення $S_{t-1}^{(1)}$ при $t = 1$ (див. нижче підрозділ "Вибір початкових умов").

Покажемо, як залежать ЕС першого порядку (формула (79)), обрахована на момент часу t, від раніш отриманих згладжених величин:

$$
\begin{aligned}
S_t^{(1)} &= \alpha z_t + \beta S_{t-1}^{(1)} = \alpha z_t + \beta[\alpha z_{t-1} + \beta S_{t-2}^{(1)}] = \\
&= \alpha z_t + \alpha\beta z_{t-1} + \beta^2[\alpha z_{t-2} + \beta S_{t-3}^{(1)}] = \\
&= \alpha z_t + \alpha\beta z_{t-1} + \alpha\beta^2 z_{t-2} + \ldots + \alpha\beta^i z_{t-i} + \ldots + \beta^i z_0 = \\
&= \alpha \sum_{i=0}^{t-1} \beta^i z_{t-i} + \beta^t z_0 \ .
\end{aligned}
$$

Як видно, $S_t^{(1)}$ є лінійною комбінацією всіх попередніх рівнів, ваги яких зменшуються в геометричній прогресії. Вага рівня, віддаленого на p відліків від поточного моменту t, дорівнює $\alpha(1-\alpha)^p = \alpha\beta^p$. Наприклад, якщо постійна згладжування $\alpha = 0{,}3$, то для моменту часу:

t	рівень буде мати вагу	$0{,}3$,
$t-1$	-"-	$0{,}3(1-0{,}3) = 0{,}21$,
$t-2$	-"-	$0{,}3(1-0{,}3)^2 = 0{,}147$,
$t-3$	-"-	$0{,}3(1-0{,}3)^3 = 0{,}1029$ тощо.

ЕС другого порядку $S_t^{(2)}$ розраховують за вибіркою ЕС першого порядку $S_t^{(1)}$ за формулою:

$$S_t^{(2)} = \alpha S_t^{(1)} + \beta S_{t-1}^{(2)}, \qquad (80)$$

аналогічно *ЕС третього порядку* – за вибіркою ЕС другого порядку:

$$S_t^{(3)} = \alpha S_t^{(2)} + \beta S_{t-1}^{(3)}. \qquad (81)$$

Для практичного обчислення ЕС можна скористатися вбудованою можливістю Excel (Сервіс → Аналіз даних→ Експоненціальне згладжування).

9.3. Розрахунок оцінок параметрів і прогнозних значень

Обрахувавши експоненціальні середні різних порядків, далі обчислюються оцінки параметрів у виразі (75), та прогнозні значення рівнів ряду. Співвідношення, за якими виконуються обчислення, залежать від того, поліном якого ступеня використовується у локальній моделі часового ряду. Зазначимо, що ступінь k апроксимуючого поліному (75) звичайно не перевищує 2.

У довільний момент часу t прогноз P_{t+L} на $L = 1,2,3,\ldots$ відліків уперед обчислюється (як це випливає з виразу (76)) за формулою:

$$P_{t+L} = a_{ot} + a_{1t}L + \frac{1}{2!}a_{2t}L^2 + ... + \frac{1}{k!}a_{kt}L^k. \qquad (82)$$

Для поліному степеня $k = 0$, тобто для простої середньої, ця формула трансформується у вираз:

$$P_{t+l} = a_{0t} = S_t^{(1)}. \qquad (83)$$

Прогноз часового ряду на L точок вперед лінійним поліномом ($k = 1$) розраховується за формулою:

$$P_{t+L} = a_{0t} + a_{1t}L\ , \qquad (84)$$

де

$$a_{ot} = 2S_t^{(1)} - S_t^{(2)}, \qquad a_{1t} = \frac{\alpha}{\beta}(S_t^{(1)} - S_t^{(2)}). \qquad (85)$$

Квадратична поліноміальна модель ($k = 2$) для прогнозування має вигляд:

$$P_{t+L} = a_{0t} + a_{1t}L + \frac{1}{2}a_{2t}L^2\ , \qquad (86)$$

де

$$\left.\begin{aligned} a_{ot} &= 3(S_t^{(1)} - S_t^{(2)}) + S_t^{(3)}; \\ a_{1t} &= \frac{\alpha}{2\beta^2}[(6-5\alpha)S_t^{(1)} - 2(5-4\alpha)S_t^{(2)} + (4-3\alpha)S_t^{(3)}]; \\ a_{2t} &= \frac{\alpha^2}{\beta^2}(S_t^{(1)} - 2S_t^{(2)}) + S_t^{(3)}). \end{aligned}\right\}. \qquad (87)$$

9.4. Вибір початкових умов

Для розрахунку ЕС необхідно задати початкову умову $S_{t-1}^{(1)}$ при $t = 1$. Єдиного підходу до завдання початкових наближень немає, часто їх задають відповідно до змісту даних чи умов дослідження. Наприклад, як початкове наближення обирають $S_{t-1}^{(1)} = z_1$, особливо, коли аналізується ряд спостережень з флуктуаціями (біржовий курс тощо). Інколи обирають середнє значення, особливо коли даний ряд коливається біля свого середнього (продаж товару, попит на який стабільний у часі тощо): $S_{t-1}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} z_i$. Крім того, як початкові наближення можна обирати середні декількох початкових значень.

Якщо використовувати початкові умови першого варіанта, то при

$t=1$ згладжене значення попереднього етапу (тобто початкове наближення) $S_{t-1}^{(1)}=S_0^{(1)}$ приймається рівним першому значенню вихідного ряду z_1, тоді:

$$S_1^{(1)}=\alpha z_1+\beta S_0^{(1)}=\alpha z_1+(1-\alpha)z_1=z_1. \qquad (88)$$

Для наступних моментів часу використовують формулу (79), наприклад:

$$\text{при}\quad t=2,\quad S_2^{(1)}=\alpha z_2+\beta S_1^{(1)},$$
$$t=3,\quad S_3^{(1)}=\alpha z_3+\beta S_2^{(1)},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$t=n,\quad S_n^{(1)}=\alpha z_n+\beta S_{n-1}^{(1)}.$$

9.5. Вибір оптимальних параметра згладжування α та степеня полінома k методом ретропрогнозу

Другою проблемою при прогнозуванні методами експоненціального згладжування є вибір параметра згладжування α. Слід врахувати, що значення параметра суттєво впливає на результат прогнозу. Якщо значення α близько до одиниці, то при прогнозуванні враховують в основному вплив останніх рівнів ряду; якщо α близьке до нуля, то ваги, за якими зважуються рівні ряду, спадають повільно, що дає можливість врахувати практично всі попередні значення ряду.

Виконуючи прогноз, слід уточнити степінь k поліноміальної моделі та параметр згладжування α. Для кожних обраних значень α та k будуються моделі прогнозу на $L=1,2,3$ відліки та обчислюють прогноз на останніх q рівнях ряду, тобто виконують ретропрогноз (якщо обсяг вихідних даних n дозволяє, бажано брати $q\cong 20\div 30$). Для цих рівнів розраховують відхилення значень ретропрогнозу за локальною моделлю $P_{t+L}(k,\alpha)$ від фактичних даних z_{t+L} часового ряду та дисперсії цих відхилень за формулою:

$$Q(k,\alpha,L)=\frac{1}{q}\sum_{t=n-q+1}^{n}[z_{t+L}-P_{t+L}(k,\alpha)]^2. \qquad (89)$$

Уточнення виконують шляхом зіставлення дисперсій відповідних прогнозів для кожного параметра згладжування α, степеня поліному k при прогнозі на $L=1,2,3$ точки.

Двійка значень k^*,α^*, для яких буде отримано мінімум показника $Q(k,\alpha,L)$ (окремо для кожного інтервалу прогнозу L), вважаються

оптимальними й використовуються для подальших розрахунків прогнозу за локальною моделлю $P_{t+L}(k^*, \alpha^*)$.

10. Особливості згладжування часових рядів в присутності аномальних даних. Усунення аномальних даних

10.1. Виявлення аномальних даних в часових рядах

При попередньому аналізі вихідних даних іноді можна спостерігати окремі значення, які суттєво відрізняються від значень, що знаходяться поруч. Це так звані аномальні дані, що виникли, наприклад, через випадкові збої у вимірювальній або обчислювальній техніці. Аномальні дані (АД) можуть зустрічатися у будь-яких експериментальних даних: як у вибіркових сукупностях, так і у часових рядах, причому у кожному з цих випадків не вилучені АД стають причиною появи серйозних помилок у результатах , отриманих з використанням цих даних.

Для розробки методів виявлення АД в часовому ряді (послідовності) $\{y_i\}$, $i = \overline{1,n}$ принциповим є питання сталості, незмінності в часі імовірнісних характеристик процесу $y(t)$. Якщо це має місце, процес $y(t)$ називається *стаціонарним* у вузькому сенсі. Для такого процесу розподіл ймовірностей його значень лишається однаковим і постійним для будь-якого моменту часу t. За цих умов до виявлення АД серед елементів часового ряду $\{y_i\}$, $i = \overline{1,n}$ можна залучати методи виявлення АД у вибіркових сукупностях, моделлю яких є множина значень, отриманих у серії випробувань випадкової величини Y. Однак зазначені методи не враховують існуючих між елементами часового ряду кореляційних зв'язків, наявність яких дозволяє зробити пошук та виокремлення АД більш успішним.

При наявності відомостей про автокореляційну функцію $\rho(t)$ процесу $y(t)$ доцільно спершу від вихідного часового ряду перейти до різницевого ряду $\{r_i\}, i = \overline{1,n-1}$, де

$$r_i = y_{i+1} - y_i \ . \qquad (90)$$

За відсутності АД математичне сподівання такого ряду дорівнює нулю, а дисперсія може стати меншою за σ_y^2:

$$\sigma_r^2 = M\{(y_{i+1} - y_i)^2\} = M\{y_{i+1}^2 + y_i^2 - 2y_{i+1}y_i\} = 2\sigma_y^2[1 - \rho(\Delta_t)], \quad (91)$$

тобто за сильної автокореляції для виявлення АД більш сприятливі умови будуть при обробці різницевого ряду. Слід також враховувати, що поряд з можливим зменшенням дисперсії σ_r^2 відносно σ_y^2 при наявності АД після переходу до різницевого ряду виникає своєрідне "підсилення" аномальності за рахунок розмноження АД. Наприклад, якщо елемент y_{i+1} є аномальним, аномальність (різного знаку) будуть мати вже дві різниці: r_i та r_{i+1}.

Якщо ряд $\{y_i\}$ нестаціонарний, розв'язок задачі пошуку та виявлення АД суттєво ускладнюється і стає залежним від характеру нестаціонарності.

Коли ряд нестаціонарний за математичним сподіванням і структуру його елемента можна представити у вигляді:

$$y_i = tr_i + e_i, \qquad (92)$$

є можливість шляхом згладжування вихідного ряду оцінити трендову складову, обчислити нев'язки ε_i і далі, аналізуючи послідовність нев'язок, виявити АД. Звичайно уживані методи виділення тренду шляхом поліноміальної апроксимації вихідного ряду $\{y_i\}$ або лінійних згладжуючих фільтрів при наявності у вихідній послідовності АД приводять до надто суттєвих зміщень в отриманих оцінках тренду. Через це перспективи подальшої роботи з цими оцінками вельми суперечливі, тому для оцінювання тренду слід застосовувати робастні алгоритми, серед яких найбільш поширеним є медіанне згладжування.

10.2. Метод медіанного згладжування

Застосовування методу медіанного згладжування для виявлення та усунення АД реалізується шляхом аналізу даних в межах ковзкого інтервалу, що покроково переміщується вздовж часового ряду в заданому напрямі, звичайно за наростанням індексів членів ряду. Довільному положенню ковзкого інтервалу відповідає фрагмент вихідного часового ряду з $2m+1$ елементів: $y_{i-m},\ldots,y_i,\ldots,y_{i+m}$. Згладжена оцінка $\tilde{y}_i$ визначається для середнього елемента інтервалу шляхом побудови варіаційного ряду з його елементів $y_{(1)},\ldots,y_{(m+1)},\ldots y_{(2m+1)}$ та виділення його медіани:

$$\tilde{y}_i = Med\{y_{i-m},\ldots,y_i,\ldots,y_{i+m}\} = y_{(m+1)}. \qquad (93)$$

Після цього зі складу ковзкого інтервалу виключається крайній лівий елемент та дописується справа черговий елемент часового ряду. При цьому середина інтервалу зміщується на один елемент вправо. Для нового положення ковзкого інтервалу повторюється процедура обчислення згладженої оцінки середнього елемента. У такий спосіб ковзкий інтервал

проходить повздовж усього часового ряду, починаючи з його лівого краю.

Після згладжування вихідного ряду можна оцінити трендову складову часового ряду і далі, аналізуючи послідовність нев'язок, виявити АД.

Якщо у вихідній вибірці y_i відсутні АД кратності $q > m$ (кратність q – кількість послідовних елементів ряду, що являють собою аномальності), то в якості згладжених оцінок елементів часового ряду будуть виступати тільки кондиційні елементи ряду, які не належать до аномалій. Таким чином, якщо у вихідному часовому ряді присутні АД, кратність яких не перевищує m, в згладженому ряді їх не буде, проте в часовому ряді, утвореному нев'язками, будуть присутні відповідні АД з вихідного ряду.

З точки зору забезпечення надійного робастного згладжування, стійкого до АД високої кратності q, слід використовувати ковзкий інтервал з великим $m \geq q$. З іншого боку, із ростом m, за рахунок так званої динамічної похибки медіанного фільтра, починають зростати амплітуди нев'язки для кондиційних даних, що стає завадою для надійного виділення АД на фоні збільшених (зрослих) значень нев'язок.

Одним з рішень в цьому випадку є подвійне медіанне згладжування, в якому послідовність згладжених даних $\{y_i'\}$, отриманих на першому етапі обробки вихідного часового ряду (з достатньо великим m), ще раз «підгладжується» медіанним фільтром з $m_2 < m_1$. Прикладом такої обробки є так звана *процедура „Тьюки 53"*, де $m_1 = 2$, $m_2 = 1$. Отримавши після другого згладжування ряд $\{y_i''\}$, фінальну згладжену оцінку рекомендується розраховувати за формулою:

$$\tilde{y}_i = 0{,}5y_i'' + 0{,}25\left(y_{i-1}'' + y_{i+1}''\right). \qquad (94)$$

Більш складні випадки нестаціонарності часового ряду $\{y_i\}$ вимагають застосування більш складних методів обробки даних.

ЧАСТИНА 3. ЛАБОРАТОРНІ РОБОТИ

№1. Параметрична ідентифікація моделей

Мета роботи: вирішення задачі параметричної ідентифікації, застосування вбудованих функцій Excel для аналізу статистичних даних.

Ключові поняття: ідентифікація моделі, оцінка параметрів, статистики для оцінки якості моделі.

Питання для самоконтролю

1. Визначте поняття: математична модель, ідентифікація математичної моделі.
2. Назвіть та поясніть зміст етапів ідентифікації моделі.
3. Які властивості характеризують випадкову складову в залежній змінній?
4. Чи містять випадкову складову незалежні змінні?
5. Для чого використовують функцію ЛИНЕЙН?
6. Для якого класу моделей застосовують цю функцію?
7. Який метод оцінювання параметрів закладено у цю функцію?
8. Назвіть аргументи функції ЛИНЕЙН. В чому полягають особливості завдання аргументів?
9. Які додаткові статистики обраховує функція ЛИНЕЙН?

Зміст та порядок виконання

1. Залежність $Y = f(X)$ представлена квадратичною параболою
$$Y = a_0 + a_1 X + a_2 X^2 \qquad (-1-)$$
2. Самостійно сформувати тестовий приклад, задав коефіцієнти a_0, a_1, a_2 для рівняння (-1-) у вигляді довільних констант.
3. Заповнити таблицю з чотирма колонками (Y, X, X^2, Z): задати 10 значень $X = \overline{1,10}$, розрахувати Y за тестовим прикладом (див. п.2). До залежної змінної Y додати шумову компоненту e_i та отримані дані Z занести до таблиці: $z_i = y_i + e_i$, $i = \overline{1,10}$.
4. Розраховані дані прийняти за вихідні.
5. За допомогою статистичної функції ЛИНЕЙН розрахувати для квадратичної параболи оцінки параметрів a_0, a_1, a_2.
6. Скористатися підказкою у вікні функції ЛИНЕЙН для визначення статистик, які розраховує функція.
7. Визначити розрахункові (модельні) значення $\tilde{Y}$ за моделлю (-1-), для

чого використати оцінки параметрів, отримані за допомогою функції ЛИНЕЙН.

8. Порівняти оцінки параметрів із заданими коефіцієнтами (див. п.2).
9. Побудувати графіки Z та $\tilde{Y}$, використовуючи Майстер діаграм.
10. Змінити значення Z у таблиці вихідних даних. Проаналізувати зміни в статистиках функції ЛИНЕЙН.
11. Розрахувати суму квадратів нев'язок (див. підказку функції Excel)

$$S_{ost} = \sum_{i=1}^{10} (z_i - \tilde{y}_i)^2 . \qquad (-2-)$$

12. Розрахувати коефіцієнт детермінації (див. підказку функції Excel)

$$R^2 = 1 - \frac{\sum_{i=1}^{10} (z_i - \tilde{y}_i)^2}{\sum_{i=1}^{10} (z_i - \bar{z})^2} , \qquad (-3-)$$

де $\bar{z}$ – середнє арифметичне значення залежної змінної Z.

13. Порівняти розрахункові значення S_{ost} та R^2 з відповідними статистиками функції ЛИНЕЙН.
14. Зробити висновки.

№2. Аналіз даних методом парної регресії

Мета роботи: ознайомитись з загальними відомостями про регресійні моделі та проблему ідентифікації.

Ключові поняття: регресія, парна регресія, оцінки параметрів, метод найменших квадратів (МНК), коефіцієнт детермінації, статистика Фішера.

Питання для самоконтролю

1. Що таке регресія, регресійний аналіз, парна регресія, лінійна, нелінійна регресія?
2. Для чого використовують МНК?
3. Назвіть критерій, на якому базується МНК.
4. Які Ви знаєте властивості МНК?
5. Які особливості МНК-оцінок?
6. Зміст перевірки моделі на адекватность.
7. Для чого використовується статистика Фішера?
8. Для чого використовується коефіцієнт детермінації?

Зміст та порядок виконання

1. Вихідні дані Z та X – векторні величини, представлені парами значень $\{z_i, x_i\}, i = \overline{1,n}$, між якими існує парна лінійна залежність

$$\tilde{y}_i = a_0 + a_1 x_i \ . \qquad \text{(-1-)}$$

2. Розробити макет таблиці, для поточних розрахунків.
3. Розрахувати оцінки параметрів a_1, a_0 парної лінійної регресії за формулами:

$$a_1 = \frac{n\sum_{i=1}^{n}(z_i x_i) - \sum_{i=1}^{n}(z_i)\sum_{i=1}^{n}(x_i)}{n\sum_{i=1}^{n}(x_i)^2 - (\sum_{i=1}^{n} x_i)^2}, \qquad \text{(-2-)}$$

$$a_0 = \frac{1}{n}\sum_{i=1}^{n}(z_i) - a_1\frac{1}{n}\sum_{i=1}^{n}(x_i). \qquad \text{(-3-)}$$

4. Розрахувати статистику Фішера:

$$F = \frac{n-m}{m-1}\ \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{z})^2}{\sum_{i=1}^{n}(z_i - \tilde{y}_i)^2}, \qquad \text{(-4-)}$$

5. Розрахувати коефіцієнт детермінації R^2 :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}\ , \qquad \text{(-5-)}$$

6. Розрахувати середньо-квадратичне відхилення:

$$\sigma = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}(z_i - \tilde{y}_i)^2} \qquad \text{(-6-)}$$

7. За допомогою вбудованої функції ЛИНЕЙН визначити оцінки парної лінійної регресії та додаткові статистики.
8. Порівняти відповідні статистики функції ЛИНЕЙН з результатами розрахунків за пунктами 3-6.
9. За допомогою Майстра діаграм побудувати на одному графіку вихідні $Z(X)$ та модельні $\tilde{Y}(X)$ значення.
10. Зробити висновки.

№3. Метод крокової регресії

Мета роботи: вибір структури регресійної моделі методом крокової регресії.

Ключові поняття: мультиколінеарність, кореляція, метод крокової регресії, регресори, ступінь вільності, оцінка якості моделі.

Питання для самоконтролю

1. Що таке мультиколінеарність?
2. Які наслідки мультиколінеарності?
3. Які ознаки наявності мультиколінеарності?
4. Які Ви знаєте методи усунення мультиколінеарності?
5. Що таке ступінь вільності.
6. Використання таблиць критичних значень статистик.
7. Що таке фактори та регресори моделі?
8. Охарактеризуйте етапи крокової регресії.
9. Для чого необхідно проводити комплексний аналіз якості моделі?

Порядок виконання

1. Обстежити вихідні дані на присутність колінеарних незалежних факторів. При виявленні мультиколінеарності необхідно позбутися однієї змінної в парах, що мають високі коефіцієнти парної кореляції.
2. Розрахувати на базі незалежних факторів x_1, x_2, x_3, x_4 множину регресорів:

   x_1, x_2, x_3, x_4, x_1^2, x_2^2, x_3^2, x_4^2, x_1x_2, x_1x_3, x_1x_4, x_2x_3, x_2x_4, x_3x_4.
3. За допомогою статистичної функції КОРРЕЛ(массив_1, массив_2) знайти коефіцієнти кореляції залежної змінної z з усіма регресорами.

Таблиця кореляцій

	x_1	x_2	x_3	x_4	x_1^2	...	x_2x_4	x_3x_4	max
z									
$abs(z)$							max		
$z-\tilde{y}^{(1)}$									
$abs(z-\tilde{y}^{(1)})$			max						
$z-\tilde{y}^{(2)}$									
...									

4. Вибрати за допомогою статистичної функції МАКС регресор, що має максимальне за модулем значення коефіцієнта парної кореляції із залежною змінною z. Включити цей регресор у модель

$$\tilde{y}^{(1)} = \tilde{a}_0 + \tilde{a}_1 x_2 x_4. \qquad \text{(-1-)}$$

5. Методом найменших квадратів за допомогою статистичної функції ЛИНЕЙН знайти оцінки коефіцієнтів a_0, a_1.

6. Розрахувати модельні значення $\tilde{y}^{(1)}$ за формулою (-1-), підставивши отримані за допомогою функції ЛИНЕЙН оцінки коефіцієнтів.

7. Знайти нев'язку ($z - \tilde{y}^{(1)}$) і визначити коефіцієнти кореляції нев'язки з усіма регресорами. Занести дані в таблицю кореляцій.

8. Вибрати регресор (див. таблицю), що має максимальне за модулем значення коефіцієнта парної кореляції із змінною ($z - \tilde{y}^{(1)}$). Включити цей регресор в ускладнену модель

$$\tilde{y}^{(2)} = \tilde{a}_0 + \tilde{a}_1 x_2 x_4 + \tilde{a}_2 x_3. \qquad \text{(-2-)}$$

9. Для нової ускладненої моделі (-2-) за допомогою МНК розрахувати оцінки коефіцієнтів a_0, a_1, a_2.

10. Розрахувати для двох моделей $\tilde{y}^{(1)}$ і $\tilde{y}^{(2)}$ F-статистику :

$$F = \frac{k\,(S^{(l)} - S^{(l+1)})}{S^{(l+1)}}, \qquad \text{(-3-)}$$

де

$S^{(l)}$ – сума квадратів нев'язок для простої моделі l-го кроку,

$S^{(l+1)}$ – сума квадратів нев'язок для ускладненої моделі наступного $(l+1)$-го кроку,

k – ступінь вільності для моделі наступного $(l+1)$-го кроку.

Значення k, $S^{(l)}$ та $S^{(l+1)}$ можна знайти у додатковій статистиці функції ЛИНЕЙН.

12. Перевірити умову $F > F_{кр}(p, k1, k2)$, виконання якої свідчить про доцільність ускладнення моделі, що суттєво збільшило точність апроксимації моделлю вихідних даних. Якщо ця нерівність не виконується, процедура крокової регресії переривається, і в якості моделі для подальшої роботи використовується більш проста модель, тобто модель, отримана не на поточному, а на попередньому кроці підбору елементів структури моделі.

13. Як тільки процес крокового ускладнення моделі виявиться неефективним, слід виконати за t-критерієм Стьюдента

$$t_j = \frac{\left|\tilde{a}_j\right|}{\tilde{\sigma}\{\tilde{a}_j\}}, \quad j = \overline{0, p}, \qquad \text{(-4-)}$$

перевірку значущості оцінок параметрів моделі, знайдених на попередніх кроках. Для значущих параметрів моделі обов'язкова умова $t_j > t_{кр}(p,k)$. Якщо будуть виявлені незначущі оцінки, для яких $t_j < t_{кр}(p,k)$, то треба виключити їх з моделі, перевіривши точність спрощеної моделі за F-статистикою.

Табличні значення $F_{кр}(p,k1,k2)$ та $t_{кр}(p,k)$, де p – довірча ймовірність, можна знайти у додатку.

14. П.п..3–12 повторюються, поки крокове ускладнення моделі буде ефективним.
15. Якщо виникають сумніви щодо ускладнення або спрощення моделі, провадиться комплексний аналіз якості моделі за F-, t-статистиками, за показниками коефіцієнта детермінації R^2, за значеннями стандартної похибки оцінювання залежної змінної z. Значну частину з цієї інформації можна отримати з додаткової статистики, яку повертає функція ЛИНЕЙН.
16. Побудувати графіки залежної змінної z та кращої моделі $\tilde{y}$.

№4. Нелінійна множинна регресія

Мета роботи: аналіз задач нелінійної множинної регресії.

Ключові поняття: нелінійна регресія, перетворення до лінійної форми моделей, логарифмування, зворотне перетворення.

Питання для самоконтролю

1. Особливості нелінійних регресій.
2. В чому полягає різниця простих та множинних нелінійних регресій?
3. Навіщо нелінійні регресії перетворювати до лінійної форми?
4. Які види функцій допускають перетворення до лінійної?
5. Способи перетворення нелінійних моделей у лінійні

Порядок виконання

1.Залежність $y = f(x_1, x_2, x_3)$ представлена у вигляді трьох нелінійних регресій:

експоненціальної – $y_1 = a_0 e^{a_1 x_1} e^{a_2 x_2} e^{a_3 x_3}$;

степеневої – $y_2 = a_0 x_1^{a_1} x_2^{a_2} x_3^{a_3}$; (-1-)

показникової – $y_3 = a_0 a_1^{x_1} a_2^{x_2} a_3^{x_3}$.

2.Для приведення моделей до лінійного вигляду прологарифмувати кожну з них:

$$\ln y_1 = \ln a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3,$$
$$\ln y_2 = \ln a_0 + a_1 \ln x_1 + a_2 \ln x_2 + a_3 \ln x_3, \quad (-2-)$$
$$\ln y_3 = \ln a_0 + x_1 \ln a_1 + x_2 \ln a_2 + x_3 \ln a_3.$$

3.Зробити заміни

$$z_1 = \ln y_1, \quad z_2 = \ln y_2, \quad z_3 = \ln y_3,$$
$$x_1' = \ln x_1, \quad x'_2 = \ln x_2, \quad x_3' = \ln x\ , \quad (-3-)$$
$$a_0' = \ln a_0, \quad a_1' = \ln a_1, \quad a'_2 = \ln a_2, \quad a_3' = \ln a_3.$$

4.Переписати моделі (-1-) з урахуванням замін (-3-):

$$z_1 = a_0' + a_1 x_1 + a_2 x_2 + a_3 x_3\ ,$$
$$z_2 = a_0' + a_1 x_1' + a_2 x_2' + a_3 x_3', \quad (-4-)$$
$$z_3 = a_0' + a_1' x_1 + a_2' x_2 + a_3' x_3.$$

5.Знайти оцінки коефіцієнтів $\tilde{a}_j, j = \overline{0,3}$ за допомогою статистичної функції ЛИНЕЙН.

6.Якщо це необхідно, виконати зворотне перерахування отриманих оцінок a'_j у коефіцієнти вихідних моделей (-1-):

$$a_0 = \exp(a_0'),\ a_1 = \exp(a_1'),\ a_2 = \exp(a_2'),\ a_3 = \exp(a_3'). \quad (-5-)$$

Одержимо вектор шуканих коефіцієнтів для кожної із трьох моделей.

7.Для кожної з вихідної моделі (-1-) отримати розрахункові значення залежної змінної $\tilde{y}_i,\ i = \overline{1,n}$.

8.Проаналізувати якість моделей. Вибрати кращу модель, використовуючи в якості критеріїв всі відомі статистики.

9.Оцінити за *t*-критерієм Стьюдента значимість оцінок параметрів для кращої моделі.

№5. Виділення тренда та прогнозування часового ряду. Статистика Дарбіна-Уотсона.

Мета роботи: виділення тренду та прогнозування часового ряду з використанням можливостей Excel.

Ключові поняття: часовий ряд, тренд, компоненти часового ряду, види функцій тренду.

Питання для самоконтролю

1. Що таке часовий ряд?
2. Що таке екстраполяція?
3. Назвіть компоненти часового ряду, визначити їх сутність.
4. Які Ви знаєте види функцій тренду?
5. Для чого використовують статистику Дарбіна-Уотсона?
6. В якому діапазоні знаходяться значення цієї статистики?
7. Доведіть, що граничні значення статистики дорівнюють саме цим значенням.

Порядок виконання

1. Вибрати свій варіант вихідних даних, представлених часовим рядом z_t, $t = \overline{1,n}$.
2. Побудувати за допомогою *Майстра діаграм* графік вихідних даних.
3. У поле діаграми викликати контекстне меню для елемента *Ряд даних*, обрати команду *Додати лінію тренду*.
4. У вікні *Лінія тренду* вибрати *Лінійний* вид тренду.
5. На вкладці *Параметри* виправити назву лінії тренду та відмітити:
   ☑ Показати рівняння тренду;
   ☑ Додати коефіцієнт детермінації R^2.
6. Використовуючи параметри рівняння тренду (на діаграмі), розрахувати модельні значення для лінійного тренду.
7. На цю ж діаграму додати *Логарифмічний*, *Степеневий* та *Експоненціальний* види тренду.
8. Побудувати новий графік вихідного часового ряду й додати на одну діаграму три поліноміальні тренди для степені поліному $p = 2; 4; 6$
9. Для трьох поліномів на вкладці *Параметри* відмітити *Прогноз наперед* на 3 періоди.
10. Пояснити різницю результатів прогнозу, отриманих за допомогою різних видів поліноміальних трендів.
11. Побудувати новий графік вихідного часового ряду й додати на одну діаграму три тренди *Лінійна фільтрація* на 2, 4 и 10 точок.
12. Розрахувати для кожного виду тренду ряд модельних значень.
13. Розрахувати для кожного виду тренду суму квадратів нев'язок.
14. За допомогою функції ЛИНЕЙН оцінити коефіцієнти a_i для тренду 4-го степеня:
$$\tilde{y} = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 \ .$$
15. Розрахувати суму квадратів нев'язок для цієї моделі.
16. Вибрати кращий тип тренду для апроксимації вихідних даних z_t за коефіцієнтом детермінації R^2 і за сумою квадратів нев'язок.

17. Застосувати статистику Дарбіна-Уотсона для виявлення автокореляції в значеннях нев'язки.
18. Вибрати кращий вид тренду за статисткою Дарбіна-Уотсона.
19. Зробити висновки.

№6. Оцінка точності прогнозу за ретроданими

Мета роботи: оцінка точності прогнозу, отриманого на базі моделі, параметри якої обчислені методом найменших квадратів.

Ключові поняття: ретродані, точність прогнозу, оцінка точності прогнозу.

Питання для самоконтролю

1. Яка методика прогнозування за ретроданими?
2. Особливості використання функцій Excel ЛИНЕЙН та ТЕНДЕНЦИЯ?
3. Як впливають особливості вихідних даних на точність прогнозу?
4. Як Ви вважаєте, переускладнена або проста модель забезпечить більш точний прогноз?
5. Як оцінити точність прогнозу за ретроданими?

Порядок виконання

1. Для прогнозу часового ряду z_t, $t=\overline{1,n}$, об'єм вихідної вибірки $n=40$, використати два рівняння тренду:
   а) модель $\tilde{y}^{(1)}$, підібрану за коефіцієнтом детермінації R^2 (*Майстер діаграм* → *Додати лінію тренду* → тип *Поліноміальний тренд* → вкладка *Параметри*), значущість параметрів моделі $\tilde{y}^{(1)}$ перевірити за *t*-статистикою Стьюдента;
   б) – переускладнену модель $\tilde{y}^{(2)}$, степінь поліному якої вище чим у моделі $\tilde{y}^{(1)}$ на 2, тобто $p^{(2)}=p^{(1)}+2$;
   – або простішу модель $\tilde{y}^{(2)}$, степінь поліному якої нижче чим у моделі $\tilde{y}^{(1)}$, наприклад на 2, тобто $p^{(2)}=p^{(1)}-2$.
2. Оцінити параметри моделей $\tilde{y}^{(1)}$ та $\tilde{y}^{(2)}$ за допомогою методу найменших квадратів (функція ЛИНЕЙН).
3. Найти для трьох точок k=41,42,43 прогнозні значення за двома моделями, використовуючи знайдені оцінки параметрів.
4. Використовуючи скорочену вихідну вибірку ретроданих, отримати за

двома моделями прогнозні значення для трьох послідовних точок, що вже існують.

5. Для цього покроково подовжуючи часовий ряд з 30-ти до 40-ка точок, тобто почергово задаючи k=30,31,...,40, за допомогою функції ТЕНДЕНЦІЯ(*Y, X, 3_прогнозні_ значення, константа*) розрахувати три прогнозні значення в точках (k+1), (k+2), (k+3).
6. Розрахувати нев'язки e_{i1}, e_{i2}, e_{i3} для трьох точок прогнозу на всіх i кроках (див. п.5).
7. Побудувати (в одному масштабі по вісі Y) графіки нев'язок на усіх i кроках для першої, другої та третьої точок прогнозу, отриманих для двох рівнянь тренду $\tilde{y}^{(1)}$ и $\tilde{y}^{(2)}$.
8. Обчислити середній квадрат помилки прогнозу для першої, другої й третьої точок прогнозу за отриманими на кожному кроці значенням нев'язок e_i:

$$\sigma_1^2 = \frac{1}{10}\sum_{i=31}^{40} e_{i1}^2, \qquad \sigma_2^2 = \frac{1}{9}\sum_{i=32}^{40} e_{i2}^2, \qquad \sigma_3^2 = \frac{1}{8}\sum_{i=33}^{40} e_{i3}^2.$$

9. Обчислити середньоквадратичне відхилення $\sigma_1, \sigma_2, \sigma_3$ для трьох точок прогнозу.
10. За допомогою *Майстра діаграм* побудувати графік вихідного ряду z_t, $t = 40$, добавити два рівняння тренду $\tilde{y}^{(1)}$ та $\tilde{y}^{(2)}$ з трьома точками прогнозу.
11. Нанести на графіки інтервали $\pm 2\sigma$ для візуального представлення точності отриманого прогнозу за моделями.
12. Порівняти результати прогнозу вихідного ряду за двома моделями тренду.

№7. Згладжування часових рядів за допомогою зваженої ковзкої середньої. Застосування критеріїв Дарбіна-Уотсона та Аббе.

Мета роботи: згладжування часового ряду за допомогою зваженої ковзкої середньої.

Ключові поняття: зважена ковзка середня, ковзкий інтервал, критерії Дарбіна-Уотсона, Аббе.

Питання для самоконтролю

1. Що таке зважена ковзка середня? Яким чином обчислюються ваги для

ковзкої середньої?

2. Якими параметрами характеризується ковзкий інтервал?
3. Особливості застосування критерію Дарбіна-Уотсона.
4. Особливості застосування критерію Аббе.
5. В чому різниця критеріїв Дарбіна-Уотсона та Аббе?
6. Як оцінити якість згладжування часових рядів ковзким середнім?

Порядок виконання

1. Вихідні дані: z_i – часовий ряд, де $i=\overline{1,n}$, $n=40$, z_i^c – той самий часовий ряд, але з циклічною компонентою.
2. Виділити тренди часових рядів y_i и y_i^c методом крокової регресії.
3. Побудувати графіки вихідних даних и трендів часових рядів.
4. Протестувати (див. табл.1*) отримані трендові моделі на наявність автокореляції за допомогою d -статистики Дарбіна-Уотсона:
$$d=\sum_{i=1}^{n-1}(\varepsilon_{i+1}-\varepsilon_i)^2 \Big/ \sum_{i=1}^{n}\varepsilon_i^2 .$$
5. Згладжування часового ряду z_i^c, що містить циклічну компоненту, виконати методом зваженого ковзкого середнього, використовуючи поліноміальну модель 1-го, 2-го и 4-го порядку ($p=1,2,4$) для апроксимації даних на ковзкому інтервалі довжиною $k=5,7,9,11,13$ точок (див п.п. 6-7).
6. Згладжене значення в середній точці $\tilde{y}_l$ ковзкого інтервалу дорівнює середньозваженому значенню точок вихідного ряду в межах цього інтервалу:

$$\tilde{y}_l=(y_1*w_1+y_2*w_2+...+y_i*w_i+...+y_k*w_k)/\sum w_i ,$$

де w_i – вагові коефіцієнти, $i=\overline{1,k}$,

$\sum w_i$ – сума вагових коефіцієнтів.

Для степені поліному $p=1$ ваги $w_i=1$, $i=\overline{1,k}$.

Для $p=2,\ 4$ ваги і суми $\sum w_i$ взяти з таблиць 2* и 3* відповідно.

7. Виконати серії згладжувань за наступними параметрами:
   а) p=1, k=3,5,7,9;
   б) p=2, k=5,7,9,11;
   в) p=4, k=7,9,11,13.
8. Побудувати три діаграми для степенів поліному p=1,2,4 для всіх серій згладжувань (див. п. 7).
9. Перевірити результати згладжувань на наявність корельованих

нев'язок ε_i за допомогою критерію Аббе:

$$A = \sum_{i=1}^{n-1} (\varepsilon_{i+1} - \varepsilon_i)^2 \Big/ 2 \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 .$$

10. Вибрати за критерієм Аббе кращі моделі для степенів поліному p=1,2,4.
11. Проаналізувати можливість оцінювання якості згладжування даних за середньою сумою квадратів нев'язок.

Таблиця 1*. Граничні значення d_1 та d_2 $(n = 40,\ p$=0,95).

p=1		p=2		p=3		p=4		p=5	
d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
1,44	1,54	1,39	1,6	1,34	1,66	1,29	1,72	1,23	1,79

Таблиця 2*. Вагові коефіцієнти для згладжування (p=2).

k	$\sum w_i$	Вагові коефіцієнти w_i										
		1	2	3	4	5	6	7	8	9	10	11
11	429	-36	9	44	69	84	**89**	84	69	44	9	-36
9	231	-21	14	39	54	**59**	54	39	14	-21		
7	21	-2	3	6	**7**	6	3	-2				
5	35	-3	12	**17**	12	-3						

Таблиця 3*. Вагові коефіцієнти для згладжування (p=4).

k	$\sum w_i$	Вагові коефіцієнти w_i												
		1	2	3	4	5	6	7	8	9	10	11	12	13
13	2431	110	-198	-135	110	390	600	**677**	600	390	110	-135	-198	110
11	429	18	-45	-10	60	120	**148**	120	60	-10	-45	18		
9	429	15	-55	30	135	**179**	135	30	-55	15				
7	231	5	-30	75	**131**	75	-30	5						

№8. Усунення аномальних даних в часових рядах

Мета роботи: згладжування часового ряду за допомогою медіанного згладжування, виявлення та усунення аномальних даних (АД) з часового ряду, застосування процедури „Тьюки 53".

Ключові поняття: часовий ряд, інтервал згладжування, медіанне згладжування, аномальні дані, процедура „Тьюки 53".

Питання для самоконтролю

1. Що таке аномальні дані?
2. Що таке кратність АД?
3. В чому полягає сенс процедури Тьюки 53?
4. Що таке медіанне згладжування?
5. Що таке динамічна помилка медіанного згладжування?

Порядок виконання

Вихідний часовий ряд z_i, $i=\overline{1,n}$, $n=40$, можливо містить АД.

1. Для аналізу даних с неправдоподібними значеннями використати процедуру Тьюки 53:
   1.1. Із ряду $\{z_i\}$ сформувати новий згладжений ряд $\{\tilde{z}_i\}$. Довжина ковзкого інтервалу l=5. Визначити медіанне середнє для першого ковзкого інтервалу $z_1,...,z_5$ і записати його в згладженій (новій) послідовності елементом $\tilde{z}_3$. Для кожного наступного положення ковзкого інтервалу розрахувати медіанні значення, з яких одержати ряд $\{\tilde{z}_i\}$, що буде коротше вихідного ряду на чотири елементи.
   1.2. Із ряду $\{\tilde{z}_i\}$ сформувати послідовність $\{\tilde{\tilde{z}}_i\}$. Довжина ковзкого інтервалу l=3, медіана вибирається із трійок послідовних значень.
   1.3. Оцінити значення випадкової складової $\{e_i\}$, для чого поелементно відняти від вихідного ряду $\{z_i\}$ ряд $\{\tilde{\tilde{z}}_i\}$: $\varepsilon_i = z_i - \tilde{\tilde{z}}_i$, $i=\overline{4,37}$.
2. Побудувати графіки:
   а) часових рядів $\{z_i\}$, $\{\tilde{z}_i\}$ та $\{\tilde{\tilde{z}}_i\}$;
   б) нев'язок $\{\varepsilon_i\}$.
3. Розрахувати середнє арифметичне та середньоквадратичне відхилення (СКВ) для нев'язок $\{\varepsilon_i\}$.
4. Побудувати гістограму розподілу нев'язок, оцінивши розкид значень і задавши границі інтервалів самостійно:
   Сервис → Анализ данных → Гистограмма
   (При відсутності цієї можливості в Excel виконати наступні дії: *Сервис → Надстройки → ☑ Пакет анализа.*)
5. Знайти ряд модулів нев'язок $\{\varepsilon_i\}$.
6. Получити ряд $\{\varepsilon_i^{ранж}\}$, упорядкувавши за зростанням модулі нев'язок $\{\varepsilon_i\}$ із збереженням номера кожного *i*-го значення.
7. Побудувати графік $\{\varepsilon_i^{ранж}\}$.
8. Розрахувати середнє квадратичне відхилення (СКВ) для ряду $\{\varepsilon_i^{ранж}\}$,

виключаючи послідовно *k* останніх значень ряду: k = 10, 9, 8,...,1, 0.

9. Показати динаміку зміни одинадцяти значень СКВ на графіку.
10. Зробити висновки про наявність і положення АД у вихідному ряді $\{z_i\}$.

№9. Метод гармонічних вагових коефіцієнтів

Мета роботи: прогнозування часових рядів методом гармонічних вагових коефіцієнтів.

Ключові поняття: метод гармонічних вагових коефіцієнтів, гармонічні ваги.

Питання для самоконтролю

1. Для чого призначений метод гармонічних вагових коефіцієнтів?
2. Особливість цього методу?
3. Як розраховуються вагові коефіцієнти і в чому сенс такого розрахунку?

Порядок виконання

1. Вихідні дані представлені часовим рядом z_i, $i=\overline{1,n}$.
2. За допомогою крокової регресії виділити тренд ряду $\tilde{y}_i$, для якого регресорами є час t у степенях $t, t^2, ..., t^6$, наприклад:

$$\tilde{y} = a_0 + a_1 t + a_2 t^2 + ... + a_6 t^6. \qquad (-1-)$$

3. Розрахувати прирости для трьох прогнозних точок $p = 3$:

$$\begin{aligned} &\omega_{t+1} = \tilde{y}_{t+1} - \tilde{y}_t, \quad t=\overline{1,n-1}, \\ &\omega_{t+2} = \tilde{y}_{t+2} - \tilde{y}_t, \quad t=\overline{1,n-2}, \\ &\omega_{t+3} = \tilde{y}_{t+3} - \tilde{y}_t, \quad t=\overline{1,n-3}. \end{aligned} \qquad (-2-)$$

4. Розрахувати ваги m_{t+1} (побудувати ряд гармонічних ваг):

$$m_{t+L} = \sum_{j=L}^{n-1} \frac{1}{n-j}, \quad t=\overline{1,n-L}, \quad L=1,2,3, \quad \text{де } m_L = 0. \qquad (-3-)$$

5. Знайти коефіцієнти:

$$\begin{aligned} &c^1_{t+1}, \quad t=\overline{1,n-1}: \quad c^1_2 = \frac{m_2}{n-1}, ..., \quad c^1_{t+1} = \frac{m_{t+1}}{n-1}, \qquad \sum_{t=1}^{n-1} c^1_{t+1} = 1; \\ &c^2_{t+2}, \quad t=\overline{1,n-2}: \quad c^2_3 = \frac{m_3}{n-2}, ..., \quad c^2_{t+2} = \frac{m_{t+2}}{n-2}; \end{aligned} \qquad (-4-)$$

c_{t+L}^{L}, $t=\overline{1,n-L}$: $c_{L+1}^{L}=\dfrac{m_{t+L}}{n-L}$.

6. Розрахувати середньозважені прирости (для розрахунку можна використати функцію Excel СУММПРОИЗВ(массив_1, массив_2)):

$$\begin{aligned}&\overline{\omega}_{t+1}=\sum_{t=1}^{n-1}c_{t+1}^{1}\omega_{t+1},\\ &\ldots\\ &\overline{\omega}_{t+L}=\sum_{t=1}^{n-L}c_{t+L}^{L}\omega_{t+L}.\end{aligned} \qquad \text{(-5-)}$$

7. Розрахувати прогнозні значення для трьох точок:

1-й варіант	2-й варіант	3-й варіант
$\tilde{y}_{n+1}=\tilde{y}_n+\overline{\omega}_{t+1}$		
$\tilde{y}_{n+2}=\tilde{y}_{n+1}+\overline{\omega}_{t+1}$	$\tilde{y}_{n+2}=\tilde{y}_n+\overline{\omega}_{t+2}$	
$\tilde{y}_{n+3}=\tilde{y}_{n+2}+\overline{\omega}_{t+1}$	$\tilde{y}_{n+3}=\tilde{y}_{n+1}+\overline{\omega}_{t+2}$	$\tilde{y}_{n+3}=\tilde{y}_n+\overline{\omega}_{t+3}$

(-6-)

8. Побудувати графіки вихідного ряду, тренду з трьома прогнозними значеннями.

№10. Метод авторегресійних моделей

Мета роботи: прогнозування часових рядів методом авторегресійних моделей.

Ключові поняття: ідентифікація математичної моделі, авторегресія, корелограма.

Питання для самоконтролю

1. Для чого призначений метод авторегресійних моделей?
2. Особливості цього методу?
3. Що таке авторегресійна модель?
4. Які дані використовують для формування авторегресійної моделі?
5. Що таке автокореляція?
6. Корелограма: побудова та її використання.
7. Як визначається порядок p авторегресійної моделі?
8. Як перевірити ефективність введення корекції тренду методом авторегресійної моделі.

Порядок виконання

1. Записати вихідні дані z_i, $i = 1,2,...,n$ та розрахувати множину регресорів: $t, t^2, ..., t^6$, де t – час .
2. Визначити трендові модельні значення $\tilde{y}_i$, використовуючи оцінки коефіцієнтів, отримані за допомогою функції ЛИНЕЙН.
3. Визначити відхилення (нев'язки) моделі $\varepsilon_i = z_i - \tilde{y}_i$, їх квадрати та дисперсію відхилення D_ε для 40 та 10 рівнів.
4. Розрахувати автокореляційну функцію r_i, $i = \overline{1,10}$ та межі довірчого інтервалу $-1/n \pm 2/\sqrt{n}$. Визначити, чи потрапляють значення r_i в довірчий інтервал.
5. Побудувати корелограму-1 за даними п.4.
6. Визначити орієнтовний порядок авторегресійної моделі *p*.
7. Визначити орієнтовну (розширену) структуру авторегресійної моделі та сформувати вихідні дані для розрахунку МНК-оцінок коефіцієнтів авторегресійної моделі:

$$Z = \begin{bmatrix} \varepsilon_{p+1} \\ \varepsilon_{p+2} \\ \dots \\ \varepsilon_n \end{bmatrix}; \quad X = \begin{bmatrix} \varepsilon_p & \varepsilon_{p-1} & \dots & \varepsilon_1 \\ \varepsilon_{p+1} & \varepsilon_{p+2} & \dots & \varepsilon_2 \\ \dots & \dots & \dots & \dots \\ \varepsilon_{n-1} & \varepsilon_{n-2} & \dots & \varepsilon_{n-p} \end{bmatrix}$$

8. Використовуючи методику визначення та виключення надлишкових регресорів (метод вимітання), уточнити структуру та коефіцієнти авторегресійної моделі.
9. Використовуючи отриману уточнену авторегресійну модель, обчислити поправки δ^*_{n+1}, δ^*_{n+2}, δ^*_{n+3} до прогнозу тренду:

$$\delta^*_{n+1} = a_1\varepsilon_n + a_2\varepsilon_{n-1} + ... + a_p\varepsilon_{n-p},$$

$$\delta^*_{n+2} = a_1\delta^*_{n+1} + a_2\varepsilon_n + a_3\varepsilon_{n-1} + ... + a_p\varepsilon_{n-p+1},$$

$$\delta^*_{n+3} = a_1\delta^*_{n+2} + a_2\delta^*_{n+1} + a_3\varepsilon_n + ... + a_p\varepsilon_{n-p+2},$$

де p – порядок уточненої авторегресійної моделі.

10. Ввести корегуючі поправки δ^*_{n+1}, δ^*_{n+2}, δ^*_{n+3} до відповідних прогнозних оцінок тренду:

$$y^*_{n+1} = \tilde{y}_{n+1} + \delta^*_{n+1},$$

$$y^*_{n+2} = \tilde{y}_{n+2} + \delta^*_{n+2},$$

$$y^*_{n+3} = \tilde{y}_{n+3} + \delta^*_{n+3}.$$

11. Побудувати графіки вихідних даних z_i, тренду $\tilde{y}_i$ та трьох точок

прогнозу y^*_{n+1}, y^*_{n+2}, y^*_{n+3}.

12. Розрахувати уточнені рівні тренду y^*_{i+j}, $i=\overline{p,n-j}$, j=1,2,3 для всього часового ряду. Для цього, використовуючи отриману авторегресійну модель, обрахувати поправки δ^*_{i+1}, δ^*_{i+2}, δ^*_{i+3} до тренду:

$$\delta^*_{i+1} = a_1\varepsilon_i + a_2\varepsilon_{i-1} + ... + a_p\varepsilon_{i-p}, \qquad i=\overline{p,n-1},$$

$$\delta^*_{i+2} = a_1\delta^*_{i+1} + a_2\varepsilon_i + a_3\varepsilon_{i-1} + ... + a_p\varepsilon_{i-p+1}, \quad i=\overline{p,n-2},$$

$$\delta^*_{i+3} = a_1\delta^*_{i+2} + a_2\delta^*_{i+1} + a_3\varepsilon_i + ... + a_p\varepsilon_{i-p+2}, \quad i=\overline{p,n-3},$$

де p – порядок уточненої авторегресійної моделі.

13. Ввести корегуючі поправки δ^*_{i+j}, $i=\overline{p,n-j}$, j=1,2,3 до відповідних оцінок ретропрогнозу тренду:

$$y^*_{i+1} = \tilde{y}_{i+1} + \delta^*_{i+1},$$
$$y^*_{i+2} = \tilde{y}_{i+2} + \delta^*_{i+2},$$
$$y^*_{i+3} = \tilde{y}_{i+3} + \delta^*_{i+3}.$$

14. Побудувати графіки нев'язки $\{\varepsilon_i\}$ та поправок $\{\delta^*_{i+j}\}$, $i=\overline{p,n-j}$, j=1,2,3.

15. За результатами отриманих у п.11 скорегованих даних ретропрогнозу оцінити ефективність побудованої авторегресійної моделі:

а) розрахувати оцінки випадкового компоненту $\varepsilon^*_i = z_i - y^*_i$;

б) розрахувати квадрати випадкового компоненту $(\varepsilon^*_i)^2$;

в) визначити дисперсію D_ε випадкового компонента 1) за всією вибіркою $\{\varepsilon^*_i\}$, 2) на останніх 10 точках цієї вибірки;

г) порівняти дисперсію помилки D_ε та дисперсію випадкового компонента D_{ε^*}.

16. Побудувати графіки: 1)вихідного ряду $\{z_i\}$; 2)тренду $\{\tilde{y}_i\}$; 3)випадкового компоненту $\{\varepsilon^*_i\}$; 4)інформативного сигналу = тренд + поправка за авторегресійною моделлю: $\{y^*_i\}$.

17. Розрахувати автокореляційну функцію, r_i, $i=\overline{1,10}$ для випадкового компоненту ε^*_i.

18. Визначити, чи потрапляють нові значення r_i у довірчий інтервал.

19. Побудувати корелограму-2 та оцінити ефективність уточнення тренду часового ряду.

№11. Метод експоненціального згладжування

Мета роботи: прогнозування часових рядів методом експоненціального згладжування

Ключові поняття: експоненціальні середні різних порядків, параметри експоненціального згладжування.

Питання для самоконтролю

1. Для чого призначений метод експоненціального згладжування?
2. Особливість методу експоненціального згладжування.
3. Як обраховуються експоненціальні середні різних порядків?
4. Вибір початкових умов експоненціального згладжування.
5. Вибір параметру згладжування та степеня поліному.

Порядок виконання

1. Вибрати свій варіант вихідних даних, представлених часовим рядом z_t, $t=\overline{1,n}$.
2. Задати параметр згладжування α, $\alpha \in [0{,}03 \div 0{,}4]$.
3. Розрахувати фактор затухання $\beta = 1-\alpha$.
4. Розрахувати експоненціальні середні 1-го, 2-го, 3-го порядку.
5. Розрахувати оцінки коефіцієнтів a_{0t}, a_{1t} для лінійної моделі експоненціального середнього та оцінки ретропрогнозу на 1, 2, 3 відліки вперед ($l=1,2,3$).
6. Розрахувати оцінки коефіцієнтів a_{0t}, a_{1t}, a_{2t} для квадратичної моделі та оцінки ретропрогнозу на 1, 2, 3 відліки вперед ($l=1,2,3$).
7. Знайти, починаючи з одинадцятого рівня часового ряду, квадрати відхилень фактичних рівнів часового ряду від прогнозних для лінійної моделі. Розрахувати відповідні дисперсії.
8. Знайти квадрати відхилень для квадратичної моделі. Розрахувати відповідні дисперсії.
9. Скопіювати всі розрахунки для заданого параметру згладжування α. В копії виправити значення α і відповідні формули в таблиці.
10. Розрахувати за наведеною схемою декілька таблиць з різними параметрами згладжування α.
11. Розраховані дисперсії для всіх точок прогнозу для різних α звести в таблицю.
12. Визначити оптимальні параметри лінійної та квадратичної моделей для

кожного інтервалу упередження та обрахувати відповідні прогнози для часового ряду.

13. За допомогою Майстра діаграм побудувати графіки залежності дисперсії від параметрів згладжування α для однієї, двох та трьох точок прогнозу.

ЛІТЕРАТУРА

1. Френкель А.А. Прогнозирование производительности труда: методы и модели. – М.: Экономика, 1989. – 214 с.
2. Тюрин Ю.И., Макаров А.А. Статистический анализ данных на компьютере/ Под ред. В.Э. Фигурнова.. – М.: ИНФРА-М, 1998. – 528 с.
3. Лук'яненко І.Г., Краснікова Л.І. Економетрика: Підручник. – К.: Товариство "Знання", КОО, 1998. – 494 с.
4. Кендэл М. Временные ряды. – М.: Финансы и статистика, 1981. – 199 с.
5. Льюнг Л. Идентификация систем. – М.: Наука, 1991. – 432с.
6. БСЭ, т.16. – М.: Сов. энциклопедия, 1974. – 616с.
7. Растригин Л.А., Маджаров Н.Е. Введение в идентификацию объектов управления. – М.: Энергия, 1977. –216с.
8. Пугачев В.С. Теория вероятностей и математическая статистика. – М.: Наука, 1979. – 496с.
9. Себер Дж. Линейный регрессионный анализ. – М.: Мир, 1980. – 456с.
10. Демиденко Е.З. Линейная и нелинейная регрессии. – М.: Финансы и статистика, 1981. – 302с.
11. Драйпер Н., Смит Г. Прикладной регрессионный анализ: В 2-х кн. Кн.2/ Пер.с англ.–2-е изд., перераб. и доп. – М.: Финансы и статистика. 1987. – 351с.: ил.
12. Айвазян С.А. и др. Прикладная статистика: Исследование зависимостей: Справ. изд.– М.: Финансы и статистика, 1985.– 487с., ил.
13. Шаракшанэ А.С., Железнов И.Г. Испытания сложных систем. Учеб. пособие для вузов. М., "Высш.школа", 1974.
14. Кузнєцов Ю.М., Скляров Р.А. Прогнозування розвитку технічних систем: Навчальний посібник. Під загальною редакцією проф. Ю.М. Кузнєцова. – К.: ТОВ "ЗМОК" –ПП "ГНОЗИС", 2004.-с.323:іл.
15. Матвієнко В.Я. Прогностика. – К.: Українські пропілеї, 2000. – 484с.
16. Снитюк В.Є. Прогнозування. Моделі. Методи. Алгоритми: Навчальний посібник. – К.: "Маклаут", 2008. – 364с.
17. Згуровский М.З. Технологическое предвидение/ М.З. Згуровский, Н.Д. Панкратова. К.: ІВЦ "Видавництво "Політехніка"", 2005. – 156 с.

18. Таблицы математической статистики. Большев Л.Н., Смирнов Н.В. – М.: Наука, Главная редакция физико-математической литературы, 1983. – 416 с.
19. Вучков И. и др. Прикладной линейный регрессионный анализ / И. Вучков, Л. Бояджиева, Е. Солаков / Пер. с болг. и предисл. Ю.П. Адлера. – М.: Финансы и статистика, 1987. – 239 с.: ил.
20. Кучин В.Л., Якушев Е.В. Управление развитием экономических систем. – М.: Экономика,1990. –157с.
21. Плотинский Ю.М. Математическое моделирование динамики социальных процессов. – М.: Изд-во Моск. ун-та, 1992. – 133с.
22. Толстова Ю.Н. Измерение в социологии. – М.: ИНФРА-М, 1998. – 224с.
23. Ядов В.А. Стратегия социологического исследования. Описание, объяснение, понимание социальной реальности. – М.: "Добросвет", 1999. – 596с.
24. Мидлтон М.Р. Анализ статистических данных с использованием Microsoft Excel для Office XP / М.Р. Мидлтон; Пер. с англ.; Под ред. Г.М. Кобелькова. – М.: БИНОМ. Лаборатория знаний, 2005. – 296с.: ил.
25. Додж М., Кината К., Стинсон К. Эффективная работа с Excel 7.0 для Windows 95/ Пер. с англ. – СПб: Питер, 1996. – 1040с.: ил.
26. Лондар С.Л. Економетрія засобами MS Excel: Навч. посіб./ С.Л. Лондар, Р.В. Юринець. –К.: Вид-во Європ. ун-ту, 2004. – 242с. – Бібліогр.: с.238.
27. Толбатов Ю.А. Econometrika: Підручник для студентів екон. спеціальн. вищих навч. закладів. – К.: Четверта хвиля, 1997. –320с.: іл.
28. Ржевський С.В. Вступ до економетрії. Навчальний посібник для студентів екон. спеціальн. – К.: Вид-во Європ. ун-ту фінансів, інформ. систем, менеджм. і бізнесу, 1999. – 93с.

ДОДАТОК

Число ступенів вільності k	t-статистика t(p,k)		F-статистика F (p,k1,k)					
	Рівень ймовірності		Рівень ймовірності P=0,05			Рівень ймовірності P=0,01		
	P=0,05	P=0,01	k1=1	k1=2	k1=3	k1=1	k1=2	k1=3
1	12,7	6,32	161,5	199,5	215,72	4052,1	4999	5403,5
2	4,3	2,92	18,51	19	19,16	98,49	99,01	99,17
3	3,18	2,35	10,13	9,55	9,28	34,12	30,81	29,46
4	2,78	2,13	7,71	6,94	6,59	21,2	18	16,69
5	2,57	2,01	6,61	5,79	5,41	16,26	13,27	12,06
6	2,45	1,94	5,99	5,14	4,76	13,74	10,92	9,78
7	2,36	1,89	5,59	4,74	4,35	12,25	9,55	8,45
8	2,31	1,86	5,32	4,46	4,07	11,26	8,65	7,59
9	2,26	1,83	5,12	4,26	3,63	10,56	8,02	6,99
10	2,23	1,81	4,96	4,1	3,71	10,04	7,56	6,55
11	2,2	1,8	4,84	3,98	3,59	9,65	7,2	6,22
12	2,18	1,78	4,75	3,88	3,49	9,33	6,93	5,95
13	2,16	1,77	4,67	3,8	3,41	3,07	6,7	5,74
14	2,14	1,76	4,6	3,74	3,34	8,86	6,51	5,56
15	2,13	1,75	4,54	3,68	3,29	8,68	6,36	5,42
16	2,12	1,75	4,49	3,63	3,24	8,53	6,23	5,29
17	2,11	1,74	4,45	3,59	3,2	8,4	6,11	5,18
18	2,1	1,73	4,41	3,55	3,16	8,28	6,01	5,09
19	2,09	1,73	4,38	3,52	3,13	8,18	5,93	5,01
20	2,09	1,73	4,35	3,49	3,1	8,1	5,85	4,94
21	2,08	1,72	4,32	3,47	3,07	8,02	5,78	4,87
22	2,07	1,72	4,3	3,44	3,05	7,94	5,72	4,87
23	2,07	1,71	4,28	3,42	3,03	7,88	5,66	4,76
24	2,06	1,71	4,26	3,4	3,01	7,82	5,61	4,72
25	2,06	1,71	4,24	3,38	2,99	7,77	5,57	4,68
26	2,06	1,71	4,22	3,37	2,98	7,72	5,53	4,64
27	2,05	1,71	4,21	3,35	2,96	7,68	5,49	4,6
28	2,05	1,7	4,2	3,34	2,95	7,64	5,45	4,57
29	2,05	1,7	4,18	3,33	2,93	7,6	5,42	4,54
30	2,04	1,7	4,17	3,32	2,92	7,56	5,39	4,51
40	2,02	1,68	4,08	3,23	2,84	7,31	5,18	4,31
60	2	1,67	4	3,15	2,76	7,08	4,98	4,13
∞	1,96	1,64	3,84	2,99	2,6	6,64	4,6	3,78