МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ СЕМЕНА КУЗНЕЦЯ

Малярець Л. М.
Ковальова К. О.

Лабораторний практикум
з навчальної дисципліни

"ЕКОНОМЕТРИКА" В СЕРЕДОВИЩІ MATLAB

Навчальний посібник

Харків. ХНЕУ ім. С. Кузнеця, 2015

УДК 330.43
ББК 65в6
М 21

Рецензенти: докт. екон. наук, професор, завідувач кафедри економічної кібернетики і прикладної економіки Харківського національного університету ім. В. Н. Каразіна *Меркулова Т. В.*; докт. екон. наук, доцент, проректор з підготовки наукових кадрів Східноєвропейського університету економіки та менеджменту м. Черкаси *Ус Г. О.*

Рекомендовано до видання рішенням вченої ради Харківського національного економічного університету імені Семена Кузнеця.

Протокол № 11 від 18.04.2015 р.

Малярець Л. М.

М 21 Лабораторний практикум з навчальної дисципліни "Економетрика" в середовищі MATLAB : навчальний посібник / Л. М. Малярець, К. О. Ковальова. – Х. : ХНЕУ ім. С. Кузнеця, 2015. – 192 с. (Укр. мов.)

ISBN 978-966-676-617-8

Подано матеріал для виконання лабораторних робіт з навчальної дисципліни, коротко викладено основні теоретичні відомості за відповідною темою, наведено приклади розв'язання типових задач, варіанти лабораторних робіт.

Рекомендовано для використання студентами у процесі неперервної математичної підготовки, магістрам, аспірантам для проведення наукових досліджень на основі побудови економетричних моделей у середовищі MATLAB, економістам-практикам для обґрунтування й ухвалення управлінських рішень з використанням комп'ютерних програм.

УДК 330.43
ББК 65в6

ISBN 978-966-676-617-8

Вступ

Сучасний фахівець з економіки повинен обробляти великі масиви даних та щоденно проводити їх аналіз, використовуючи комп'ютери та різне їх програмне забезпечення. Тому вивчення дисципліни "Економетрика" супроводжується виконанням лабораторних робіт у різних середовищах.

Одним із кращих програмних пакетів для вирішення економетричних задач є програмне середовище *MATLAB* компанії MathWorks. Її унікальність порівняно з іншими програмними продуктами (MSExcel, statistica та ін.) полягає в наявності вбудованого Toolbox – *Econometrics Toolbox*.

Econometrics Toolbox – набір інструментів моделювання на основі принципів поведінки економічних систем. *Econometrics Toolbox* містить функції для моделювання економічних процесів. Надається набір моделей, які можна калібрувати за даними і використовувати для симуляцій та прогнозування.

Econometrics Toolbox підтримує ітеративний процес ідентифікації і тестування одно- і багатовимірних фінансових та економічних моделей.

Тулбокс організовує повний цикл – від розроблення до аналізу моделі: аналіз даних і попереднє оброблення, ідентифікація моделі, оцінка параметрів, симуляція та прогнозування.

Детальний приклад того, як використовувати *Econometrics Toolbox* наведено у даному лабораторному практикумі.

Метою даного посібника є висвітлення виконання лабораторних робіт з навчальної дисципліни "Економетрика" з використанням програмного продукту *MATLAB* для формування в студентів комплексного та науково-практичного бачення методів виявлення та кількісного опису причинно-наслідкових взаємозв'язків в економіці, а також закономірностей змін економічних показників та їх систем у часі.

У процесі вивчення навчальної дисципліни "Економетрика" з використанням середовища *MATLAB* відбувається формування у студентів таких професійних компетентностей:

- аналізу й обробки даних з використанням вбудованих функцій середовища *MATLAB*, необхідних для розв'язання поставлених економічних задач;

- використання вбудованих функцій тулбоксів Statistics Toolbox й Econometrics Toolbox середовища *MATLAB* для обробки економічних даних відповідно до поставленої задачі, аналізу результатів економетричного моделювання й обґрунтування отриманих висновків;
- опису економічних процесів і явищ на основі побудови стандартних економетричних моделей, аналізу та змістовної інтерпретації отриманих результатів.

У результаті вивчення навчальної дисципліни студент набуває знання основних понять економетричного аналізу, основних методів оцінювання невідомих параметрів економетричних моделей, методів перевірки статистичних гіпотез про параметри побудованих моделей, основних методів діагностики економетричних моделей, основних функцій *MATLAB,* що використовуються в економетричному аналізі, а також уміння застосовувати стандартні методи побудови економетричних моделей, обробляти статистичну інформацію й отримувати статистично обґрунтовані висновки, робити змістовні висновки з результатів економетричного моделювання та оволодіння програмним продуктом *MATLAB*, використовуючи матричні операції, вбудовані функції *Statistics Toolbox* й *Econometrics Toolboix*, призначені для рішення економетричних задач.

Лабораторна робота 1
Однофакторний дисперсійний аналіз (ANOVA)

Мета роботи: вивчити способи дослідження взаємозв'язку даних, розбитих на групи, за допомогою дисперсійного аналізу (***An****alysis* ***of va****riance*), використовуючи вбудовані функції *MATLAB* (*Statistics Toolbox*).

Основні задачі лабораторної роботи:

1. Привести наявні дані до потрібної структури, тобто створити окремі стовпці із числовими значеннями для кожної із заданих категорій.

2. Побудувати діаграму розкиду вихідних даних.

3. Перевірити виконання необхідних умов однофакторного дисперсійного аналізу, використовуючи критерій Бартлета (функції `jbtest` і `barttest`).

4. Провести однофакторний дисперсійний аналіз із використанням вбудованої функції anova1.

5. Проаналізувати вихідні параметри функції anova1 і діаграму розмаху середніх арифметичних значень.

6. Порівняти розрахункове й табличне значення *F*-критерію, зробити висновки про статистичну значущість між середніми значеннями факторів.

7. Якщо у ході виконання ANOVA отриманий статистично значущий результат, провести попарні порівняння досліджуваних факторів для виявлення розходжень між ними, використовуючи вбудовану функцію multcompare.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

1.1. Основні поняття дисперсійного аналізу

Дисперсійний аналіз (***AN****alysis* ***Of VA****riance* – *ANOVA*) використовується для виявлення впливу на досліджуваний показник деяких факторів, що звичайно не піддаються кількісному вимірюванню, а вимірюються на номінальних або порядкових шкалах. Зазначені фактори називають незалежними змінними, а досліджуваний показник з погляду впливу на нього обраних факторів – залежна змінна (пояснювальна змінна, результівна ознака).

Основна ідея дисперсійного аналізу міститься в порівнянні "факторної дисперсії", що обумовлена впливом фактора, і "залишковою дисперсією", обумовленою випадковими причинами.

У цілому дисперсійний аналіз може бути розділений на кілька видів: одновимірний (одна залежна змінна) і багатовимірний (багато залежних змінних); однофакторний (одна групуюча змінна) і багатофакторний (декілька групуючих змінних) – з можливою взаємодією між факторами; з простими вимірюваннями (залежна змінна вимірюється лише один раз) і з повторними (залежна змінна вимірюється кілька разів).

У дисперсійному аналізі передбачається, що групи розрізняються тільки середнім рівнем результівної змінної, дані в кожній групі розподілені нормально з однаковою дисперсією. Для того щоб перевірити гіпотезу про рівність групових середніх нормальних сукупностей з однаковими дисперсіями, достатньо перевірити за критерієм Фішера нульову гіпотезу про рівність факторної і залишкової дисперсій.

З математичної статистики відомо розкладання загальної суми квадратів відхилень значень ознаки x, що спостерігаються від загальної середньої $\bar{x}$ на факторну суму квадратів відхилень групових середніх від загальної середньої, яка характеризує розсіювання між групами та залишкову суму квадратів відхилень значень групи від своєї середньої, яка характеризує розсіювання всередині груп:

$$\sum (x - \bar{x})^2 = \sum (\bar{x}_{гр} - \bar{x})^2 + \sum (x - \bar{x}_{гр})^2. \qquad (1.1)$$

У регресійному аналізі загальна сума квадратів відхилень змінної y від середнього значення $\bar{y}$ розкладається на дві частини – "пояснену" і "непояснену":

$$\sum (y - \bar{y})^2 = \sum (\hat{y}_x - \bar{y})^2 + \sum (y - \hat{y}_x)^2 \qquad (1.2)$$

загальна сума квадратів відхилень	=	сума квадратів відхилень, пояснена регресією	+	залишкова сума квадратів відхилень

або $SS_y = SS_{y_x} + SS_\varepsilon$. Цю суму квадратів розглядають як суму детермінованої і випадкової компонент.

Такий розклад на складові має число ступенів свободи:

$$df_y = df_{y_x} + df_\varepsilon, \qquad (1.3)$$

де $df_y = n-1$, оскільки на n відхилень $(y_i - \bar{y})$ накладений один зв'язок – сума всіх цих відхилень дорівнює нулю $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$;

$df_\varepsilon = n-1-m$, тут m – кількість пояснювальних змінних.

Для числа ступенів свободи обчислених значень має виконуватись:

$$df_{y_x} = df_y - df_\varepsilon = (n-1) - (n-1-m) = m. \qquad (1.4)$$

Розділивши кожну суму квадратів на відповідне число ступенів свободи, отримаємо дисперсії на одну ступінь свободи. Р. Фішер запропонував усі викладки дисперсійного аналізу оформлювати у вигляді стандартної таблиці (m – число параметрів, що оцінюють у рівнянні регресії з незалежними змінними, n – число спостережень) (табл. 1.1).

Таблиця 1.1

Зведена таблиця дисперсійного аналізу

Компоненти дисперсії	Число ступенів свободи df	Дисперсія на одну ступінь свободи	Дисперсійне відношення F	Рівень значущості
Факторна	$df_{y_x} = m$	$S^2_{y_x} = \dfrac{\sum(\hat{y}_x - \bar{y})^2}{m}$	$F_p = \dfrac{S^2_{y_x}}{S^2_\varepsilon}$	α
Залишкова	$df_\varepsilon = n-m-1$	$S^2_\varepsilon = \dfrac{\sum(y - \hat{y}_x)^2}{n-m-1}$		
Загальна	$df_y = n-1$	$S^2_y = \dfrac{\sum(y - \bar{y})^2}{n-1}$		

Якщо $F_p > F_{0,01}(df_{y_x}, df_\varepsilon)$, результівна ознака Y і фактор X не є незалежними, вони пов'язані між собою, між ними є статистична (кореляційна) залежність. Якщо $F_p < F_{0,05}(df_{y_x}, df_\varepsilon)$, результівна ознака Y і фактор X є незалежними.

1.2. Теоретичні відомості про функції Matlab, які використовуються в даній лабораторній роботі

ANOVA1 **Однофакторний дисперсійний аналіз (ANOVA)**

Синтаксис

```
p = anova1(X)
p = anova1(X,group)
p = anova1(X,group,'displayopt')
[p,table] = anova1(...)
[p,table,stats] = anova1(...)
```

Опис

`p = anova1(X)` функція дозволяє провести однофакторний дисперсійний аналіз для порівняння середніх арифметичних значень однієї або декількох вибірок однакового обсягу. Вибірки визначаються вхідним аргументом `X`. `X` задається як матриця з розмірністю `mxn`, де `m` – кількість спостережень у вибірці (кількість рядків `X`), `n` – кількість вибірок (кількість стовпців матриці `X`). Вибірки повинні бути незалежними. Вихідним аргументом функції є рівень значущості `p` нульової гіпотези. Нульова гіпотеза полягає в тому, що всі вибірки в матриці `X` узяті з однієї генеральної сукупності або з різних генеральних сукупностей з рівними середніми арифметичними. `P` є ймовірністю помилки першого роду, або ймовірністю необґрунтованого відкидання нульової гіпотези. Якщо значення $p \approx 0$, то нульова гіпотеза може бути відкинута, тобто хоча б одне середнє арифметичне відрізняється від інших значень. Дослідник вибирає критичний рівень значущості $p_{КР}$ для умови прийняття нульової гіпотези $p \geq p_{кр}$. У більшості практичних випадків $p_{КР}$ приймають рівним 0,05; 0,01.

У проведенні дисперсійного аналізу загальна дисперсія ділиться на дві складові:

міжгрупову дисперсію – дисперсію середніх арифметичних значень вибірок матриці X щодо загального середнього;

внутрішньогрупову дисперсію – дисперсію значень вибірок щодо вибіркових середніх.

Результати обчислень відображуються у двох графічних вікнах. У перше вікно виводиться таблиця з результатами однофакторного дисперсійного аналізу, у друге – діаграма розмаху для середніх арифметичних за заданими вибірками.

Таблиця з результатами однофакторного дисперсійного аналізу, що наведена на рис. 1.1, містить 6 стовпців:

1) вид дисперсії (*Source*):
внутрішньогрупова (*Columns*),
міжгрупова (*Error*),
загальна (*Total*);

2) суму квадратів різниць (*SS*) між середнім арифметичним і значеннями вибірки за кожним видом дисперсії;

3) число ступенів свободи за кожним видом дисперсії (*df*);

4) середнє значення суми квадратів різниць (*MS*) за кожним видом дисперсії: *SS/df*;

5) значення статистики Фішера (*F*-статистики) для MS;

6) значення рівня значущості *p* (*Prob>F*) для обчисленого значення статистики *F*. Якщо величина *F* збільшується, то значення *p* повинне зменшуватися.

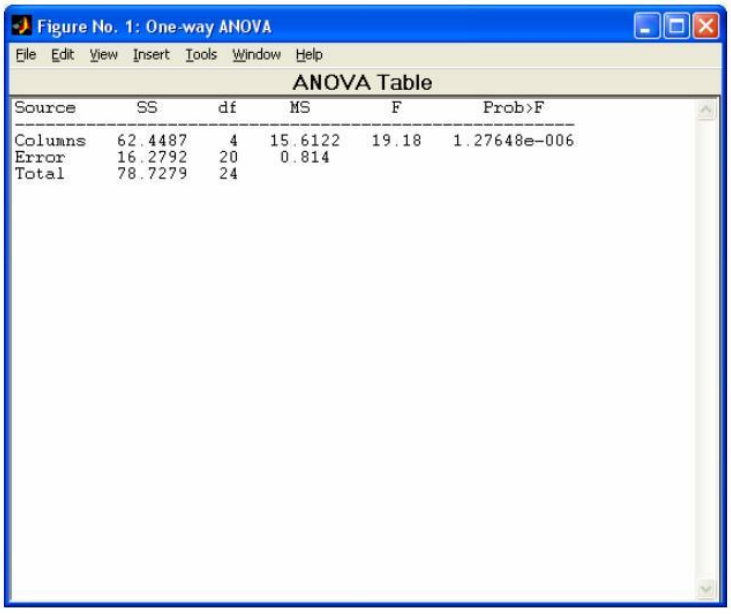


Рис. 1.1. **Діалогове вікно One-way ANOVA**

Діаграма розмаху середніх арифметичних значень будується за вибірками, тобто за стовпцями матриці *X* , і є аналогом діаграми розмаху для медіан

вибірок, визначеної за допомогою функції *boxplot*. Чим більша різниця між центральними лініями на діаграмі розмаху (середніми арифметичними вибірок), тим більша різниця між вибірковими значеннями статистики *F* і менші відповідні значення рівня значущості p.

p = anova1(*X,group*) – вхідний аргумент *group* задає мітки вибірок *X* і назв відповідних їм графіків на діаграмі розмаху. Мітки вибірок *group* задаються вектором рядків або масивом ланцюгів. Кількість елементів вектора *group* повинне дорівнювати кількості стовпців матриці *X*.

Вибірки можуть бути задані як вектор *X*. Ділення елементів X на вибірки виконується за допомогою вхідного аргументу *group*. *Group* може бути заданий вектор, символьний масив (вектор рядків) або масив ланок. Належність елемента вектора *X* до вибірки визначається однаковим значенням відповідного елемента вектора *group*. Розмірність векторів *X* і *group* повинна співпадати. Значення вектора *group* є мітками відповідних графіків вибірок на діаграмі розмаху середніх арифметичних.

Векторна форма задання вибірок *X* за згрупованою змінною *group* дозволяє проводити однофакторний дисперсійний аналіз за вибірками неоднакового обсягу. Якщо елемент вектора згрупованої змінної *group* містить порожній рядок, порожній масив середніх або нечислове значення NaN, то відповідний елемент даних у векторі *X* ігнорується при обчисленні.

p = anova1(*X,group,'displayopt'*) – вхідний аргумент *'displayopt'* дозволяє явно задати відображення графічних вікон з таблицею результатів дисперсійного аналізу та діаграмою розмаху *'displayopt'='on'* або знехтувати висновком графічних вікон *'displayopt'='off'*.

Значення за замовчуванням 'displayopt'='on'.

[*p,table*] = *anova1*(...) – функція повертає таблицю з результатами однофакторного дисперсійного аналізу в текстовій формі в командне вікно *MATLAB*.

[*p,table,stats*] = *anova1*(...) – функція повертає структуру *stats*, використовувану для проведення парних порівнянь середніх арифметичних вибірок. Перевірка параметричної гіпотези про рівність двох середніх арифметичних виконується за допомогою функції *multcompare*. Структура даних *stats* передається функції *multcompare* як вхідний аргумент.

MULTCOMPARE

Перевірка параметричних гіпотез у парному порівнянні середніх арифметичних й інших оцінок

Синтаксис

```
c = multcompare(stats)
c = multcompare(stats,alpha)
c = multcompare(stats,alpha,'displayopt')
c = multcompare(stats,alpha,'displayopt','ctype')
c = multcompare(stats,alpha,'displayopt','ctype','estimate')
c = multcompare(stats,alpha,'displayopt','ctype','estimate',dim)
[c,m] = multcompare(...)
[c,m,h] = multcompare(...)
```

Опис

c = multcompare(*stats*) – функція призначена для перевірки параметричних гіпотез у парному порівнянні середніх арифметичних або інших оцінок на основі інформації в структурі даних *stats*. Вихідними даними є матриця з результатами перевірки параметричних гіпотез. Також функція дозволяє побудувати інтерактивний графік за результатами перевірки множини параметричних гіпотез.

Під час проведення дисперсійного аналізу за множиною вибірок перевіряється нульова гіпотеза, яка полягає в тому, що всі вибіркові середні арифметичні рівні між собою. Вибірки взяті з однієї генеральної сукупності або з декількох генеральних сукупностей з однаковими значеннями середніх арифметичних. Альтернативна гіпотеза полягає в тому, що значення хоча б одного вибіркового середнього арифметичного відрізняється від інших.

Після проведення дисперсійного аналізу доцільно визначити, для якої пари вибірок середні арифметичні значення мають статистично значуще розходження. Для перевірки такої параметричної гіпотези використовується процедура множинного порівняння.

Під час перевірки простої параметричної гіпотези (нульової гіпотези) про рівність середніх однієї групи вибірок стосовно іншої за статистикою t необхідно задати рівень значущості $\alpha_{кр}$, що визначає критичне значення статистики $t_{кр}$. У більшості практичних випадків $\alpha_{кр}$ приймають рівним 0,05; 0,01. Це означає, що в 1 % або 5 % випадків буде неправильно відкинута нульова

гіпотеза. За умови збільшення груп вибірок збільшується кількість гіпотез, які перевіряють. Під час використання простої параметричної гіпотези за статистикою *t* рівень значущості $\alpha_{кр}$ буде застосовуватися до кожної гіпотези окремо, що призведе до зростання ймовірності неправильного відкидання нульової гіпотези. Процедура множинного порівняння забезпечує заданий рівень значущості для кожної перевірки.

Вихідний параметр *c* показує результати множинного порівняння у вигляді матриці з п'яти стовпців. Рядки матриці *C* відповідають результатам перевірки однієї параметричної гіпотези. Таким чином, кожен рядок *C* відповідає одній парі вибірок. Перші два значення в рядку *C* показують номери порівнюваних вибірок, третій – величину різниці середніх арифметичних порівнюваних вибірок, четвертий і п'ятий стовпці – 95 % довірчого інтервалу отриманої різниці середніх арифметичних.

Функція `multcompare` дозволяє графічно відобразити значення середніх арифметичних й їх довірчих інтервалів. Два вибіркових середніх значущо відрізняються, якщо їх довірчі інтервали не перетинаються на графіку. З накладенням меж довірчих інтервалів двох середніх арифметичних розходження між ними можна вважати статистично незначущим. За графіком можна визначити вибірки із середніми арифметичними, які значущо відрізняються від вибіркового середнього. Для цього необхідно мишею виділити графік даного вибіркового середнього. Відповідні графіки будуть виділені іншими кольорами.

`c = multcompare(stats,alpha)` – такий варіант синтаксису функції дозволяє задати рівень значущості `alpha` для розрахунку довірчих інтервалів різниць середніх арифметичних, наявних у матриці *C*. Довірча ймовірність визначається як `100(1-alpha)%`. Значення за замовчуванням `alpha = 0,05`.

`c = multcompare(stats,alpha,'displayopt')` – вхідний аргумент `'displayopt'` дозволяє відобразити графік з результатами множинного порівняння `'displayopt'='on'` або знехтувати висновок графіка `'displayopt'='off'`. Значення за замовчуванням `'displayopt'='on'`.

`c = multcompare(stats,alpha,'displayopt','ctype')` вхідний аргумент `'ctype'` дозволяє задати спосіб визначення критичних значень статистик у перевірці нульової гіпотези. Передбачено наступні значення `'ctype'` (табл. 1.2).

Таблиця 1.2

Значення вхідного аргументу `'ctype'` функції *`multcompare`*

Значення '*ctype*'	**Спосіб визначення критичних значень статистик**
'hsd', або *'tukey-kramer'*	Використовується різницевий критерій Тьюкі. Значення за замовчуванням. Критерій заснований на розподілі вибіркового розмаху за законом Стьюдента. Критерій є оптимальним для однофакторного дисперсійного аналізу й аналогічних процедур для вибірок з рівним обсягом. Доведена його обмеженість у застосуванні для однофакторного дисперсійного аналізу з різним обсягом вибірок
'lsd'	Використовується процедура визначення критичних значень статистики за найменш значущою різницею критерію Тьюкі. Ця процедура заснована на використанні простого *t* тесту. Вона застосовується, якщо попередній тест (мається на увазі *F*-статистика у однофакторному дисперсійному аналізі) показав значуще розходження порівнюваних середніх
'bonferroni'	Заснований на розрахунку критичних значень розподілу Стьюдента з коректуванням Бонферроні для використання в процедурі множинного порівняння
'dunn-sidak'	Використовуються критичні значення розраховані за розподілом Стьюдента після коректування для процедури множинного порівняння, запропонованої Даном
'scheffe'	Використовуються критичні значення розраховані за процедурою Шеффе, заснованою на розподілі Фішера. Ця процедура забезпечує однаковий рівень значущості у порівнянні лінійних комбінацій середніх

`c = multcompare(stats,alpha,'displayopt','ctype','estimate')` – вхідний аргумент *`'estimate'`* дозволяє задати вид оцінок, використовуваних у процедурі множинного порівняння. Можливі види оцінок визначаються функцією, використовуваною для розрахунку структури *`stats`*, і наведені в наступній табл. 1.3.

Таблиця 1.3

Значення вхідного аргументу *`'estimate'`* функції *`multcompare`*

Функція	**Значення '*estimate*'**
1	2
'anova1'	Значення *'estimate'* ігноруються. Завжди порівнюються середні за групами
'anova2'	*'estimate'='column'* (значення за замовчуванням) – для порівняння середніх за стовпцями, *'estimate'='row'* – для порівняння середніх за рядками

Закінчення табл. 1.3

1	2
'anovan'	Значення '*estimate*' ігноруються. Завжди порівнюються маргінальні середні генеральної сукупності, що визначаються відповідно до аргументу *dim*
'aoctool'	*'estimate'='slope', 'intercept', 'pmm'* – для порівняння значень початкового зміщення, коефіцієнтів у першому ступені незалежної змінної, маргінальних середніх генеральної сукупності. Якщо модель кореляційного аналізу не включає розділені сталі зміщення, тоді використовувати значення *'estimate'='slope'* заборонено. Також неможливо застосовувати значення *'estimate'='intercept',* якщо немає розділених коефіцієнтів у лінійному ступені незалежної змінної моделі
'friedman'	Значення ігноруються. Завжди порівнюються середні за стовпцями
'kruskalwallis'	Значення ігноруються. Завжди порівнюються середні за групами

`[c,m] = multcompare(...)` – функція повертає матрицю *m* точкових й інтервальних оцінок середніх арифметичних значень або інших порівнюваних оцінок. Перший стовпець матриці *m* містить точкові оцінки середніх арифметичних, або порівнюваних статистик, для кожної групи. Другий стовпець *m* містить величини стандартних похибок порівнюваних оцінок.

`[c,m,h] = multcompare(...)` – функція повертає покажчик *h* на графік, сформований у результаті множинного порівняння. Слід зазначити, що в рядку заголовка графіка виводиться інструкція щодо роботи з інтерактивним графіком; підписи на осі абсцис містять список груп, у яких середні арифметичні значущо відрізняються від середнього арифметичного для виділеного графіка. Для видалення заголовка графіка та підпису на осі *X* необхідно використати інтерактивні інструменти графічного вікна або команди:

```
>> title('')
>> xlabel('')
```

1.3. Розв'язування типової задачі в середовищі Matlab

Задача. Для даних, наведених у табл. 1.4, які є вартістю за різноманітні послуги, які надаються тією або іншою телекомунікаційною компанією, необхідно виконати однофакторний дисперсійний аналіз.

Таблиця 1.4

Вихідні дані задачі

Послуга	Постачальник	Ціна
Інт	NEW OTANI	**4,6**
Інт	HILTON	**5,0**
Інт	BEVERLY PLZA	**5,2**
Інт	HOL INN CONV	**5,5**
Інт	LE DUFY	**4,8**
Інт	BILTMORE	**5,1**
Інт	LE PARC	**5,7**
Інт	SHERATON GRD	**5,4**
Інт	QWERT GARD	**5,7**
Інт	PULL FILT	**5,6**
МС	HOL INN FIN	**4,3**
МС	STOUFFER	**4,4**
МС	MANDARIN	**4,9**
МС	DIVA	**5,1**
МС	GRAND HYATT	**4,5**
МС	HOLL INN GATE	**4,5**
МС	NOB HILL Int	**5,4**
МС	INN AT OPERA	**5,1**
МС	LARRIA	**5,1**
МС	SOPT Inc	**5,3**
ПД	LOMBARDY	**4,4**
ПД	SHERATON	**4,5**
ПД	HILTON	**4,9**
ПД	GRAND HYATT	**5,0**
ПД	ONE WASH CIR	**4,6**
ПД	COMDOT INN	**4,6**
ПД	LERATION INN	**5,4**
ПД	LAR IT	**5,2**
ПД	REF IT	**5,2**
ПД	BJUDI	**5,4**

Для початку роботи необхідно ввести вихідні дані в масив *X*. Привести отриманий масив *X* до потрібної структури, тобто створити окремі стовпці із числовими значеннями для кожної із заданих категорій, використовуючи навички роботи з масивами, набуті на першому курсі.

Фрагмент М-файла

```
clc
clear all
%% Ввести вихідні дані:
Int = [4.6 5.0 5.2 5.5 4.8 5.1 5.7 5.4 5.7 5.6]';
MC = [4.3 4.4 4.9 5.1 4.5 4.5 5.4 5.1 5.1 5.3]';
PD = [4.4 4.5 4.9 5.0 4.6 4.6 5.4 5.2 5.2 5.4]';
X = [Int MC PD]
group = ['In'; 'MC'; 'PD']% задамо метки вибірки X
```

Після розбивки даних за різними стовпцями з різними категоріями вони будуть виглядати так:

X =

Int	MC	PD
4.6000	4.3000	4.4000
5.0000	4.4000	4.5000
5.2000	4.9000	4.9000
5.5000	5.1000	5.0000
4.8000	4.5000	4.6000
5.1000	4.5000	4.6000
5.7000	5.4000	5.4000
5.4000	5.1000	5.2000
5.7000	5.1000	5.2000
5.6000	5.3000	5.4000

Для графічного відображення вихідних даних побудувати діаграму розкиду цін на телекомунікаційні послуги, отриманих від трьох постачальників (рис. 1.2).

Фрагмент М-файла

```
% Діаграма розкиду цін на телекомунікаційні послуги:
[n, m] = size(Int);
i = ones(n,1);
plot(i, Int,'k*', i+1, MC, 'b*',i+2,PD,'m*')
grid on
title(Діаграма розкиду цін на телекомунікаційні послуги:)
legend('Int', 'MC', 'PD')
xlabel('Номер телекомунікаційної компанії- постачальника')
ylabel('Ціна надання послуг')
```

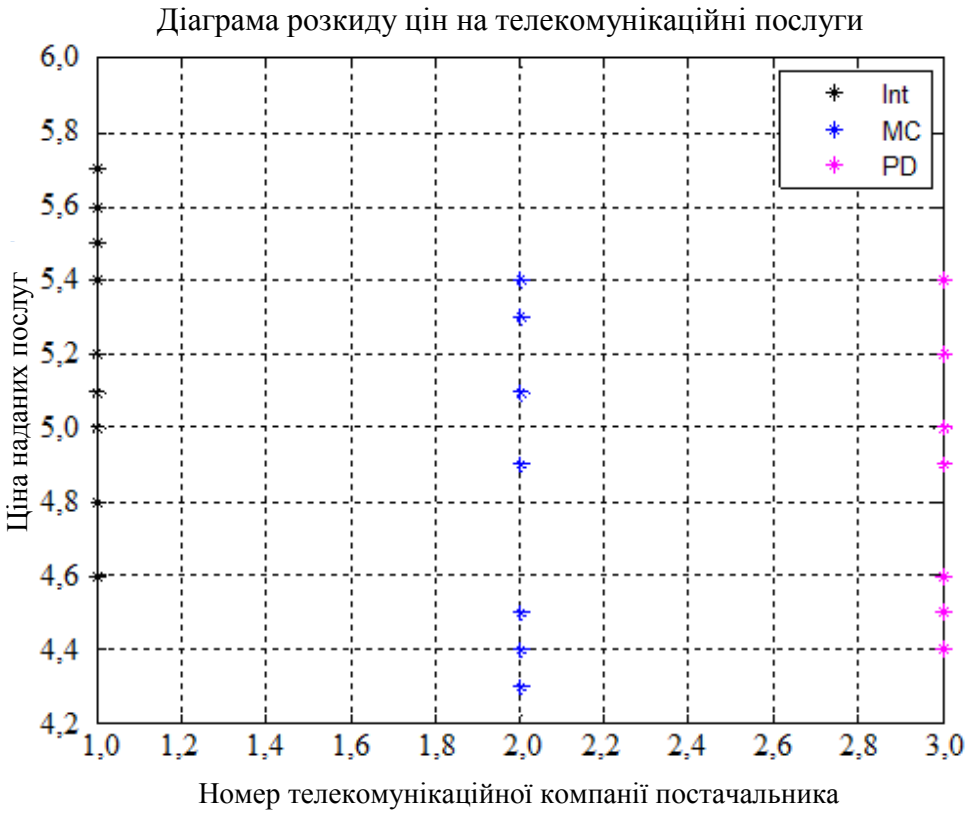


Рис. 1.2. **Діаграма розкиду цін на телекомунікаційні послуги**

Перед тим як почати однофакторний дисперсійний аналіз, варто перевірити, чи можна застосовувати цей критерій у даній ситуації. Ціни на послуги телекомунікаційних компаній – це безперервна величина, всі три групи є незалежними. Таким чином, необхідні умови 1, 2 і 5 алгоритми *ANOVA* виконуються.

Для перевірки умови нормальності розподілу в кожній із груп необхідно використати функцію `jbtest`. Функція призначена для виконання теста Яркі – Бера на несуперечність розподілу генеральної сукупності значень випадкової величини нормальному закону за вибіркою *X*. Нульова гіпотеза полягає в тому, що розподіл генеральної сукупності значень випадкової величини не суперечить нормальному закону. Нульова гіпотеза приймається, якщо $h=0$ за рівнем значущості 0,05. Якщо $h=1$, то нульова гіпотеза може бути відкинута за 0,05.

Фрагмент М-файла

```
% Оцінки параметрів нормального розподілу:
H = jbtest(Int)
H = jbtest(MC)
H = jbtest(PD)

>> H =   0
>> H =   0
>> H =   0
```

Результати теста Яркі – Бера показали, що значення у всіх трьох групах мають нормальний розподіл, тобто необхідна умова 3 про нормальний розподіл алгоритму *ANOVA* також виконується.

Для перевірки умови про рівність дисперсій досліджуваної ознаки в сукупностях, з яких відібрані вибірки, варто скористатися тестом Бартлета (функція `barttest`, що повертає `dim = 3` у випадку рівності дисперсій досліджуваних вибірок).

Фрагмент М-файла

```
%% test Bartlett
ndim = barttest(X,0.05)
```

```
>> ndim = 3
```

Оскільки `ndim = 3`, то гіпотеза H_0 про рівність дисперсій приймається, а статистика Бартлета майже підпорядковується χ^2_{m-1} – розподілу. Інакше кажучи, необхідна умова 4 про рівність дисперсій алгоритму *ANOVA* виконується.

Оскільки всі необхідні умови для застосування однофакторного дисперсійного аналізу виконані, до вихідних даних задачі можна застосувати однофакторний дисперсійний аналіз.

Для того щоб провести однофакторний дисперсійний аналіз, використовують вбудовану функцію `anova1` середовища *MATLAB*.

Фрагмент М-файла

```
group = ['In'; 'MC'; 'PD']% задамо мітки вибірки X
[p,table,stats] = anova1(X,group)
```

Результати однофакторного дисперсійного аналізу подані в таблиці `'One-way ANOVA'` (таблиця Фішера) (рис. 1.3).

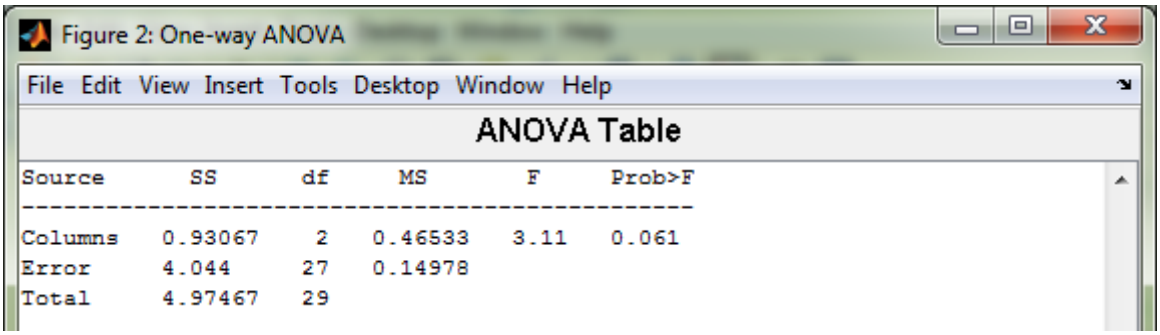


Рис. 1.3. **Діалогове вікно *'One-way ANOVA'***

У другому стовпці приводяться такі міжгрупові та внутрішньогрупові характеристики, як сума квадратів відхилень від середнього (*SS*). У третьому стовпці наведена кількість ступенів свободи (*df*), що використовується для обчислення міжгрупової і внутрішньогрупової дисперсій.

У четвертому стовпці наведена міжгрупова, внутрішньогрупова та загальна дисперсії (*MS*), визначені як відношення *SS/df* ;

Критерій F = 3,11. Досягнутий рівень статистичної значущості склав 0,061, що свідчить про існування статистично значущих розходжень між середніми значеннями в трьох порівнюваних групах.

Діаграма розмаху середніх арифметичних значень будується за вибірками. Чим більша різниця між центральними лініями на діаграмі розмаху (середніми арифметичними вибірок), тим більша різниця між вибірковими значеннями статистики F і менш відповідні значення рівня значущості p (рис. 1.4).

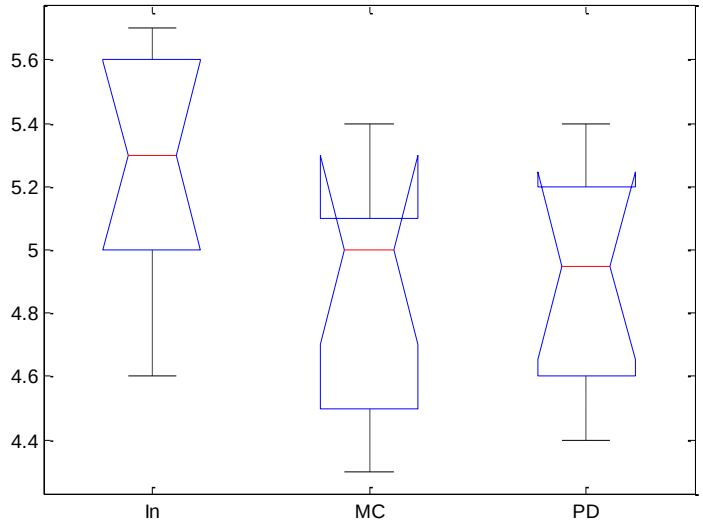
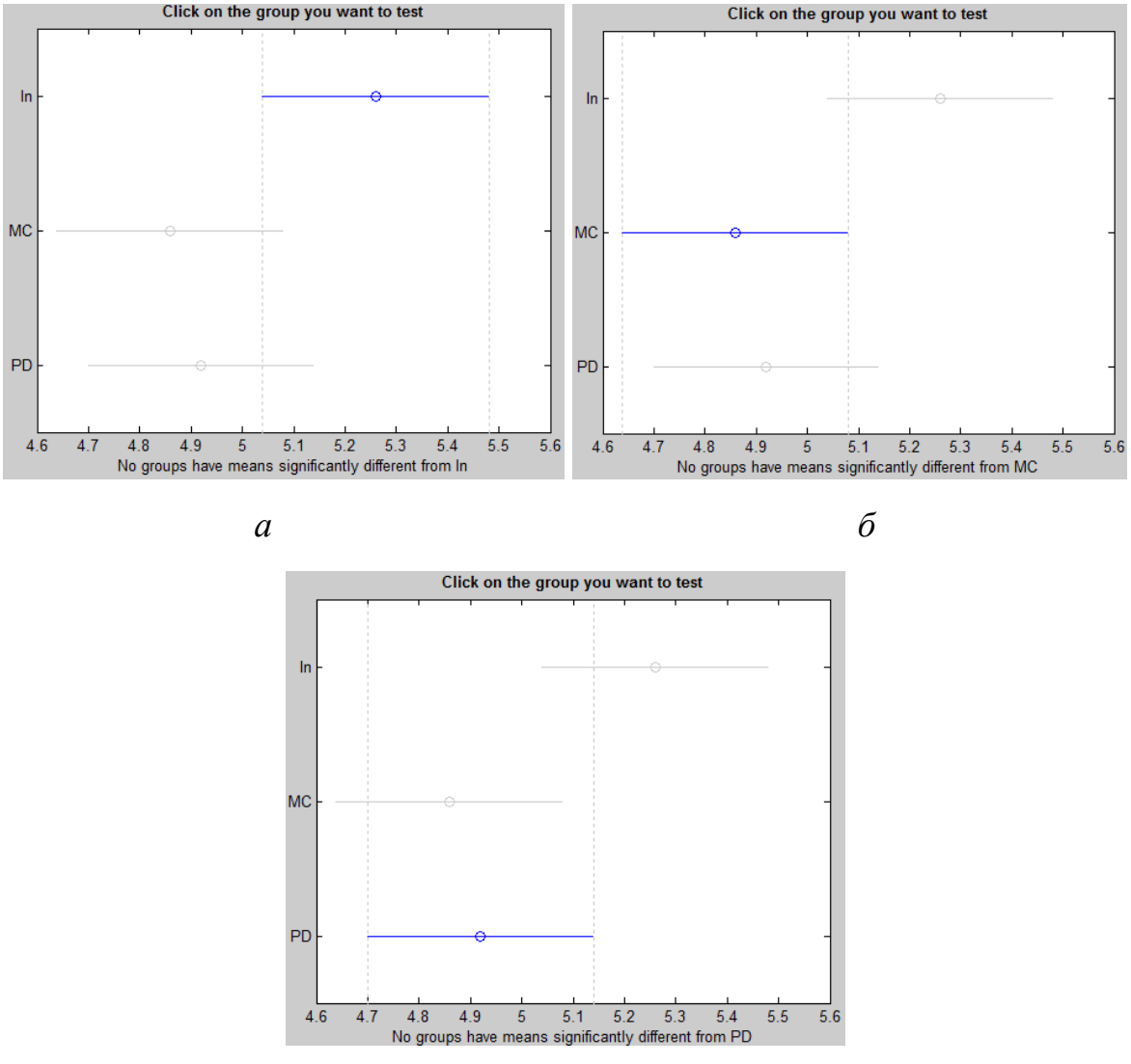


Рис. 1.4. **Діаграма розмаху середніх арифметичних значень Int, MC, PD**

Оскільки результати дисперсійного аналізу показали наявність статистично значущих розходжень між порівнюваними групами, наступним кроком є необхідність виконати апостеріорні порівняння для виявлення, між якими групами є розходження.

Оскільки під час виконання *ANOVA* отримано статистично значущий результат, слід провести попарні порівняння досліджуваних факторів для виявлення розходжень між ними, використовуючи вбудовану функцію `multcompare` (однофакторний дисперсійний аналіз для трьох вибірок з подальшим проведенням множинного порівняння оцінок середніх арифметичних, заснований на розрахунку критичних значень розподілу Стьюдента з коректуванням Бонферроні) (рис. 1.5).

Апостеріорні порівняння – це попарні порівняння досліджуваних груп для виявлення розходжень між ними. Найбільшим за розповсюдженням критерієм, що використовується для виконання попарних порівнянь, є виправлення `Bonferroni`. Результати апостеріорних порівнянь в *MATLAB* виглядають, як на рис. 1.5. У розглянутому прикладі за допомогою виправлення Бонферроні встановлено, що статистично значущих розходжень між групами не виявлено.



а *б*

в

Рис. 1.5. **Попарні порівняння досліджуваних факторів: *а – MC* і *PD, б – In* і *PD, в – In* і *MC***

Отже, аналіз завершується, тому попарні порівняння не проводяться:

```
>> c =
    1.0000    2.0000   -0.0418    0.4000    0.8418
    1.0000    3.0000   -0.1018    0.3400    0.7818
    2.0000    3.0000   -0.5018   -0.0600    0.3818
```

```
>> m =
    5.2600    0.1224
    4.8600    0.1224
    4.9200    0.1224
>> h =  2
```

Запитання для самоперевірки

1. Що таке дисперсійний аналіз?
2. Наведіть види дисперсійного аналізу.
3. Що таке однофакторний дисперсійний аналіз?
4. Запишіть математичну модель однофакторного дисперсійного аналізу.
5. Наведіть необхідні умови проведення однофакторного дисперсійного аналізу.
6. Проаналізуйте стандартну таблицю однофакторного дисперсійного аналізу, запропоновану Фішером.
7. Які функції *MATLAB* використовуються для проведення дисперсійного аналізу?

Завдання до лабораторної роботи

Задача. Три групи продавців продавали штучний товар, розфасований у різні впакування. Після закінчення строку розпродажу був зроблений тестовий контроль над випадково відібраними продавцями з кожної групи. Були отримані наступні результати, наведені в табл. 1.5 (p_1 – кількість букв у повному імені, $p = p_1/10$).

Таблиця 1.5

Вихідні дані задачі

Номер продавця	**Вид товару**	**Кількість продажів, зроблених продавцями**
1	2	3
Пр1	Плитка настінна Pierre De Buxy Sand, 25х40 см	**51** + *p*
Пр1	Плитка для підлоги Ibiza Lila, 35х35 см	**52** + *p*
Пр1	Плитка для підлоги Ibiza Aguamarina, 35х35 см	**56** + *p*
Пр1	Плитка настінна Ibiza Lila, 27х60 см	**57** + *p*
Пр2	Плитка настінна Ibiza Aguamarina, 27х60 см	**52** + *p*
Пр2	Плитка настінна Ibiza Blanco, 27х60 см	**54** + *p*
Пр2	Плитка настінна Fantasy низ, 25х35 см	**56** + *p*
Пр2	Плитка настінна Fantasy верх, 25х35 см	**58** + *p*

Закінчення табл. 1.5

1	2	3
Пр3	Плитка для підлоги Fantasy, 33х33 см	**42** + *p*
Пр3	Плитка для підлоги Reef, 33х33 см	**44** + *p*
Пр3	Плитка настінна Reef, колір голубий, 20х30 см	**50** + *p*
Пр3	Плитка настінна Reef, колір синій, 20х30 см	**52** + *p*

1. Привести наявні дані до потрібної структури, тобто створити окремі стовпці із числовими значеннями для кожної із заданих категорій.

2. Побудувати діаграму розкиду вихідних даних.

3. Перевірити виконання необхідних умов однофакторного дисперсійного аналізу, використовуючи тест Яркі – Бера та критерій Бартлета (функції `jbtest` й `barttest`).

4. Провести однофакторний дисперсійний аналіз із використанням вбудованої функції `anova1`.

5. Проаналізувати вихідні параметри функції *anova1* і діаграму розмаху середніх арифметичних значень.

6. Зрівняти розрахункове та табличне значення *F*-критерію, зробити висновки про статистичну значущість між середніми значеннями факторів.

7. Якщо виконання *ANOVA* дало статистично значущий результат, провести попарні порівняння досліджуваних факторів для виявлення розходжень між ними, використовуючи вбудовану функцію *multcompare*.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

Лабораторна робота 2
Парна лінійна регресія і кореляція

Мета роботи: набути компетентність кореляційно-регресійного аналізу за заданим значенням вибіркових даних, використовуючи вбудовані функції *Matlab* (*Statistics Toolbox*).

Основні задачі лабораторної роботи:

1. Побудувати лінійне рівняння парної регресії *y* за *x*.

2. Обчислити лінійний коефіцієнт парної кореляції, коефіцієнт детермінації та середню похибку апроксимації.

3. Оцінити статистичну значущість рівняння регресії в цілому й окремих параметрів регресії та кореляції за допомогою *F*-критерію Фішера та *t*-критерію Стьюдента.

4. Виконати прогноз заробітної плати *y* з прогнозним значенням середньодушового прожиткового мінімуму *x*, який становить 107 % від середнього рівня.

5. Обчислити довірчі інтервали для параметрів регресії, виконати точковий та інтервальний прогноз за рівнянням лінійної регресії.

6. Побудувати графік вихідних даних і рівняння регресії.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

2.1. Основні поняття кореляційно-регресійного аналізу

Мета кореляційно-регресійного аналізу полягає у встановленні факту наявності або відсутності залежностей між кількома показниками й опису цих зв'язків досить простими виразами.

Розрізняють лінійні та нелінійні регресії. Лінійна регресія знайшла широке застосування в економетриці у вигляді чіткої економічної інтерпретації її параметрів.

Лінійна парна регресія зводиться до знаходження рівняння виду: $\hat{y}_x = a + b \cdot x$ або $y = a + b \cdot x + \varepsilon$. Нелінійні регресії діляться на два класи: регресії, нелінійні щодо включених в аналіз змінних, які пояснюють, але лінійні за параметрами, які оцінюють, і регресії, нелінійні за параметрам, що оцінюють.

Параметр *b* називають коефіцієнтом регресії. Він характеризує нахил прямої до осі абсцис або $b = tg\alpha$, де α – кут, який пряма регресія утворює з віссю абсцис. Коефіцієнт регресії є мірою залежності змінної y від змінної x або мірою впливу зміною x змінної на змінну *y*. Коефіцієнт *b* вказує середню величину змінення змінної y за зміненням пояснювальної змінної x на одну одиницю, при цьому знак *b* вказує напрямок цього змінення. Якщо $b > 0$, створюється додатна лінійна регресія, що демонструє поступальний характер зміни залежної змінної за збільшенням значень пояснювальної змінної *x*. Якщо $b < 0$, то лінійна регресія буде від'ємною, тоді із збільшенням значень x значення змінної y зменшуються.

Класичний підхід до оцінювання параметрів лінійної регресії заснований на методі найменших квадратів (МНК), що дозволяє отримати такі оцінки параметрів, за яких сума квадратів відхилень фактичних значень результівної ознаки *y* від теоретичних $\hat{y}_x$ мінімальна:

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_{x_i} \right)^2 \to \min, \varepsilon_i = y_i - \hat{y}_{x_i}, \quad \sum_{i=1}^{n} \varepsilon_i^2 \to \min. \qquad (2.1)$$

Позначимо $S = \sum_{i=1}^{n} \left(y_i - \hat{y}_{x_i} \right)^2 = \sum_{i=1}^{n} \left(y_i - a - b \cdot x_i \right)^2$. Щоб знайти мінімум даної функції, треба обчислити частині похідні за кожним з параметрів a та b і прирівняти їх до нуля.

$$\frac{dS}{da} = -2\sum_{i=1}^{n} y + 2n \cdot a + 2 \cdot b\sum_{i=1}^{n} x = 0; \tag{2.2}$$

$$\frac{dS}{db} = -2\sum_{i=1}^{n} y \cdot x + 2 \cdot a\sum_{i=1}^{n} x + 2 \cdot b\sum_{i=1}^{n} x^2 = 0. \tag{2.3}$$

Отримано систему нормальних рівнянь для оцінювання параметрів a та b:

$$\begin{cases} n \cdot a + b\sum_{i=1}^{n} x = \sum_{i=1}^{n} y, \\ a\sum_{i=1}^{n} x + b\sum_{i=1}^{n} x^2 = \sum_{i=1}^{n} x \cdot y. \end{cases} \tag{2.4}$$

Методом виключення знайдено:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \; a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}. \tag{2.5}$$

Або

$$a = \bar{y} - b \cdot \bar{x}, \; b = \frac{\overline{yx} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2}, \tag{2.6}$$

де $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ і $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\overline{yx} = \frac{\sum_{i=1}^{n} y_i x_i}{n}$.

Тісноту зв'язку досліджуваних явищ оцінює лінійний коефіцієнт парної кореляції:

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \right)}}. \tag{2.7}$$

Відомо, що $-1 \le r_{xy} \le 1$. Якщо $r_{xy} = 0$, то відсутній лінійний кореляційний взаємозв'язок між результівною ознакою та фактором. Чим ближче значення коефіцієнта кореляції за модулем до 1, тим тісніший лінійний кореляційний взаємозв'язок між результівною ознакою та фактором.

Існує зв'язок між коефіцієнтом кореляції та коефіцієнтом регресії:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\overline{yx} - \overline{y}\,\overline{x}}{\sigma_x \sigma_y}. \qquad (2.8)$$

Рівняння парної лінійної регресії зручно записати в стандартизованих змінних:

$$\frac{y - \bar{y}}{\sigma_y} = r_{xy}\frac{x - \bar{x}}{\sigma_x}. \qquad (2.9)$$

Для оцінювання якості підбору парної лінійної функції використовують квадрат лінійного коефіцієнта кореляції, який називають коефіцієнтом детермінації, тобто $R^2 = r^2$.

На початковому етапі розроблення регресійної моделі перевіряється гіпотеза про наявність лінійної залежності між y та x: $H_0 : b = 0$, $H_1 : b \ne 0$. Таку гіпотезу ще називають гіпотезою про статистичну значущість коефіцієнта регресії. У разі відхилення нульової гіпотези коефіцієнт регресії вважається статистично значущим, що свідчить про наявність лінійної залежності між *y* та *x*.

Для перевірки гіпотези обчислюються *t*-значення Стьюдента: $t_b = \frac{b}{S_b}$; $t_a = \frac{a}{S_a}$ і порівнюються з табличним за числом ступенів свободи $n-2$. Якщо $t_b > t_{\alpha=0,05}(n-2)$, то гіпотезу про неістотність коефіцієнта регресії можна відхилити. Стандартні похибки S_b і S_a обчислюють за формулами:

$$S_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{S_\varepsilon^2}{\sum (x - \bar{x})^2}}; \qquad (2.10)$$

$$S_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{n-2} \cdot \frac{\sum x^2}{n\sum (x - \bar{x})^2}} = \sqrt{S_\varepsilon^2 \cdot \frac{\sum x^2}{n\sum (x - \bar{x})^2}}. \qquad (2.11)$$

Значущість лінійного коефіцієнта кореляції перевіряється на основі величини похибки коефіцієнта кореляції m_r:

$$m_r = \sqrt{\frac{1-r^2}{n-2}}. \tag{2.12}$$

Обчислене значення *t*-критерію Стьюдента знаходять за формулою:

$$t_r = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}. \tag{2.13}$$

Даний вираз свідчить, що в парній лінійній регресії $t_r^2 = F$ і $t_b^2 = F$. Отже, $t_r^2 = t_b^2$. Тобто перевірка гіпотез про значущість коефіцієнтів регресії та кореляції рівнозначна перевірці гіпотези про істотність лінійного рівняння регресії.

Знаючи значення стандартних похибок S_b і S_a, можна обчислити довірчі інтервали для параметрів рівняння регресії:

$$b \pm t_{\alpha,n-2} S_b;\ a \pm t_{\alpha,n-2} S_a. \tag{2.14}$$

За допомогою дисперсійного аналізу перевіряється значущість кореляційного зв'язку. Якщо в результаті дисперсійного аналізу виявиться, що кореляційний зв'язок не значущий, то немає сенсу проводити регресійний аналіз заданої форми, вона також буде не значущою. Для перевірки значущості кореляційного зв'язку y/x використовувати $SS_{y_x} = \eta^2 SS_y$ і $SS_\varepsilon = (1-\eta^2) SS_y$. Отримано вираз для дисперсійного відношення Фішера:

$$F_\eta = \frac{S_{y_x}^2}{S_\varepsilon^2} \cdot \frac{df_\varepsilon}{df_{y_x}} = \frac{\eta^2}{1-\eta^2} \cdot \frac{n-k}{k-1}, \tag{2.15}$$

де k – кількість інтервалів групування за x.

Обчислене значення порівняти з табличними значеннями $F_{0,01}(df_{y_x}, df_\varepsilon)$ і $F_{0,05}(df_{y_x}, df_\varepsilon)$. Якщо $F_\eta < F_{0,05}(df_{y_x}, df_\varepsilon)$, то доведена відсутність кореляційного зв'язку.

Оцінювання значущості рівняння регресії в цілому виконується за допомогою *F*-критерію Фішера. Водночас висувається нульова гіпотеза (H_0), що коефіцієнт регресії дорівнює нулю, тобто $H_0 : b = 0$.

Суму квадратів $SS_{y_x} = R^2 SS_y$ і $SS_\varepsilon = (1 - R^2) SS_y$ виразити через загальну суму квадратів SS_y і коефіцієнт детермінації R^2. Отримано вираз дисперсійного відношення Фішера:

$$F_p = \frac{SS_{y_x}}{SS_\varepsilon} \cdot \frac{df_\varepsilon}{df_{y_x}} = \frac{R^2}{1 - R^2} \cdot \frac{n - 1 - m}{m}, \tag{2.16}$$

який потрібно порівнювати з табличними значеннями: $F_{0,01}(df_{y_x}, df_\varepsilon)$ і $F_{0,05}(df_{y_x}, df_\varepsilon)$.

Значення *F*-критерію Фішера для парної лінійної регресії розраховується $F_p = \frac{r^2}{1 - r^2} \cdot (n - 2)$. Якщо $F_p > F_{0,01}(n - 2)$, то H_0 відкидається і робиться висновок про суттєвість (значущість) зв'язку, що вивчається. Якщо $F_p < F_{0,05}(n - 2)$, то H_0 не може бути відхилена без серйозного ризику зробити неправильний висновок про наявність зв'язку та рівняння регресії вважається статистично незначущим.

Одним з основних завдань побудови регресійної моделі є використання її для обчислення прогнозу. Точкове прогнозне значення $\hat{y}_x$ за $x_p = x_k$ отримують підстановкою в рівняння регресії $\hat{y}_x = a + bx$ відповідного значення x_p. Прогноз результівного показника буде більш реалістичним, якщо обчислити довірчий інтервал для значення змінної за заданим значенням змінної x:

$$\hat{y}_x - m_{\hat{y}_x} \le y_p \le \hat{y}_x + m_{\hat{y}_x}, \tag{2.17}$$

де $m_{\hat{y}_x}$ – стандартна похибка.

Формула обчислення стандартної похибки прогнозованого значення:

$$m^2_{\hat{y}_x} = \frac{S^2_\varepsilon}{n} + \frac{S^2_\varepsilon}{\sum (x - \bar{x})^2} (x - \bar{x})^2 = S^2_\varepsilon \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2} \right) \tag{2.18}$$

або

$$m_{\hat{y}_x} = S_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}}. \tag{2.19}$$

Величина стандартної похибки прогнозного середнього значення за заданим значенням $x_p = x_k$ характеризує похибку положення лінії регресії. Величина стандартної похибки $m_{\hat{y}_x}$ досягає мінімуму за $\bar{x}$ і зростає в міру віддалення від нього. Обчислення прогнозного значення $\hat{y}_x$ з 95 % довірчим інтервалом виконується за формулою:

$$\hat{y}_{x_p=x_k} \pm t_\alpha \cdot m_{\hat{y}_x}. \tag{2.20}$$

На графіку довірчі межі для $\hat{y}_x$ становлять гіперболи, розташовані по обидві сторони від лінії регресії. Фактичні значення y варіюють біля середнього значення $\hat{y}_x$. Окремі значення y можуть відхилятися від $\hat{y}_x$ на величину випадкової похибки ε, дисперсія якої є залишковою дисперсією на одну ступінь свободи. Тому помилка прогнозного індивідуального значення y включає і стандартну похибку $m_{\hat{y}_x}$, і випадкову похибку S_ε. Отже, середня похибка прогнозованого індивідуального значення $y(x_k)$ складатиме:

$$m_{y(x_k)} = S\sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x - \bar{x})^2}}. \tag{2.21}$$

Щоб мати загальне судження про якість моделі з відносних відхилень за кожним спостереженням, визначають середню похибку апроксимації як середню арифметичну просту:

$$A = \frac{1}{n}\sum \left| \frac{y - \hat{y}_x}{y} \right| \cdot 100. \tag{2.22}$$

В аналізі моделей впливу фактора на результат корисним є розрахунок коефіцієнта еластичності: $Э = f'(x)\frac{x}{y}$, де $f'(x)$ – перша похідна, яка характеризує співвідношення приростів результату і фактора для відповідної форми зв'язку. Оскільки коефіцієнт еластичності для лінійної функції не є величиною постійною, а залежить від відповідного значення *x*, то розраховується середній показник еластичності за формулою:

$$\overline{Э} = b\frac{\bar{x}}{\bar{y}}. \tag{2.23}$$

2.2. Теоретичні відомості про функції *MATLAB*, які використовуються в даній лабораторній роботі

Statistics Toolbox пропонує широкий спектр інструментів для статистичних обчислень. Основні можливості включають: регресійний аналіз та діагностику з вибором змінної, нелінійне моделювання, моделювання ймовірностей й оцінювання параметрів, аналіз чутливості з використанням генератора випадкових чисел, управління статистичними процесами та планування експерименту. Пакет включає 20 різних розподілів імовірностей, включаючи *t*, *F* та χ^2.

Для виконання лабораторної роботи з пакета ***Statistics Toolbox*** будуть потрібні функції, наведені нижче.

GLMFIT **Узагальнена лінійна регресія**

Синтаксис

```
b = glmfit(X,Y,'distr')

b = glmfit(X,Y,'distr','link','estdisp',offset,pwts,'const')

[b,dev,stats] = glmfit(...)
```

Опис

`b = glmfit(X,Y,'distr')` – функція призначена для оцінки параметрів узагальненої лінійної регресійної моделі. Вхідними аргументами функції є: `Y` – залежна змінна, `X` – матриця незалежних змінних, *'distr'* – рядкова змінна, яка визначає вид розподілу залежної змінної.

Передбачено наступні розподіли залежної змінної (табл. 2.1).

Таблиця 2.1

Значення вхідного аргументу *'distr'* функції *glmfit*

Значення '*distr*'	**Вид розподілу**
1	2
'binomial'	Біноміальне
'gamma'	Гама
'inverse gaussian'	Зворотний розподіл Гаусса

Закінчення табл. 2.1

1	2
'lognormal'	Логнормальне
'normal'	Нормальне (значення '*distr*' за замовчуванням)
'poisson'	Пуассона

Для всіх розподілів, крім біноміального, залежна змінна *Y* задається як вектор. Для біноміального закону *Y* повинна бути визначена як матриця із двома стовпцями: у першому стовпці задається число сприятливих подій, у другому – число повторних незалежних випробувань. Матриця незалежних змінних *X* повинна містити таку ж кількість рядків, що й *Y*.

Вихідна змінна *b* – вектор точкових оцінок коефіцієнтів лінійної узагальненої регресійної моделі. Цей варіант синтаксису виклику функції *glmfit* використовує канонічний взаємозв'язок параметрів розподілу з незалежними змінними.

b = glmfit(X,Y,'distr','link','estdisp',offset,pwts,'const') – вхідні аргументи *'link'*, *'estdisp'*, *offset*, *pwts*, *'const'*, призначені для управління процесом оцінювання параметрів регресійної моделі. Вхідний аргумент *'link'* задає вид взаємозв'язку між параметром розподілу μ й оцінюваною лінійною комбінацією незалежних змінних $X \cdot b$. Передбачено наступні значення *'link'* (табл. 2.2).

Таблиця 2.2

Значення вхідного аргументу *'link'* фунції *glmfit*

Значення *'link'*	**Вид залежності**	**Взаємозв'язок за замовчуванням (канонічна)**
'identity'	$\mu = xb$	*'normal'*
'log'	$\log(\mu) = xb$	*'poisson'*
'logit'	$\log(\mu/(1-\mu)) = xb$	*'binomial'*
'probit'	$\text{norminv}(\mu) = xb$	
'comploglog'	$\log(-\log(1-\mu)) = xb$	
'logloglink'	$\log(-\log(\mu)) = xb$	
'reciprocal'	$1/\mu = xb$	*'gamma'*
p (число)	$\mu^p = xb$	*'inverse gaussian'* (якщо p =-2)

Реалізація іншого виду взаємозв'язку здійснюється за допомогою *inline* або *M*-функцій. Аргумент *'link'*, що визначає вид довільного взаємозв'язку, задається як масив ланцюгів, що містить 3 елементи: 1-й елемент містить визначення функції взаємозв'язку; 2-й елемент – похідну від функції взаємозв'язку; 3-й елемент – її зворотну функцію. Нижче наведений приклад формування масиву ланцюгів *mylinks* функцій взаємозв'язку на основі *inline*-функцій:

```
FL = inline('x.^-.5')

FD = inline('-.5*x.^-1.5')

FI = inline('x.^-2')

mylinks = {FL FI FD}.
```

За використанням *M*-функцій масив осередків формується за наступним правилом:

```
mylinks = {@FL @FD @FI}.
```

Вхідний аргумент *'estdisp'* дозволяє задати величину дисперсії вибірки для біноміального закону та розподілу Пуассона. Якщо *'estdisp'='on'*, то величина дисперсії оцінюється за вибірковим даними *X*, *Y*. Для *'estdisp'='off'* точкова оцінка дисперсії приймається рівною 1. Для інших розподілів в *glmfit* значення дисперсії розраховується у всіх випадках.

Аргумент *offset* є спеціальною незалежною змінною з коефіцієнтом у регресійній моделі, дорівненим одиниці. Як приклад використання *offset* можна розглянути моделювання кількості дефектів на різноманітних поверхнях. Потрібно отримати регресійну модель, в якій залежною змінною є відносна кількість дефектів на одиницю площі поверхні. У цьому випадку кількість дефектів буде використовуватися як залежна змінна, розподілена за законом Пуассона; логарифмічна функція (*log*) буде використана для задання виду взаємозв'язку між параметром розподілу й оцінюваною лінійною комбінацією незалежних змінних (параметр `'link'`) і як параметр `offset` буде виступати вектор логарифма від площі поверхні.

Аргумент *pwts* дозволяє задати вектор вагових значень залежної змінної. Наприклад, якщо значення залежної змінної *Y(i)* є середнім арифметичним від *f(i)* вимірів, то *f* може бути використаний як вектор вагових значень.

Вхідні параметри *offset і pwts* можуть бути задані як вектори або пропущені при виклику функції. Порожні вектори *offset і pwts* будуть трактуватися як пропущені вхідні параметри. Розмірність *Y*, *offset і pwts* повинна збігатися.

Вхідний параметр *'const'* дозволяє в явному вигляді визначити, буде розраховуватися оцінка сталого члена регресійної моделі *'const'='on'* чи ні – *'const'='off'*. Значення за замовчуванням *'const'='on'*. Якщо необхідно отримати оцінку сталого члена регресійної моделі, то рекомендується використати *'const'='on'* замість задання одиничного стовпця в матриці незалежних змінних.

[b,dev,stats] = glmfit(...) – функція повертає відхилення у векторі розв'язків *dev* і структуру *stats*. Відхилення у векторі розв'язків є узагальненням залишкової суми квадратів і використовуються для порівняння ряду регресійних моделей, склад коефіцієнтів однієї з яких є підмножиною коефіцієнтів іншої. У результаті проведеного порівняння необхідно дійти статистично значущого висновку, що погрішність опису експериментальних даних регресійною моделлю з більшою кількістю коефіцієнтів менша, ніж у моделі з меншою кількістю коефіцієнтів. Тобто вибирається регресійна модель із меншою кількістю коефіцієнтів, що забезпечує мінімальну погрішність. Структура *stats* містить наступну інформацію:

stats.dfe - число ступенів свободи погрішності регресійної моделі;
stats.s - теоретична або оцінена дисперсія параметра;
stats.sfit - оцінка дисперсії параметра;
stats.estdisp - 1 - якщо оцінка дисперсії була обчислена, 0 - якщо оцінка дисперсії була задана;
stats.beta - вектор коефіцієнтів лінійної узагальненої регресійної моделі (дорівнює вихідному параметру b);
stats.se - вектор стандартних похибок коефіцієнтів лінійної узагальненої регресійної моделі b;
stats.coeffcorr - матриця коефіцієнтів кореляції коефіцієнтів b;
stats.t - значення статистики t Стьюдента для вектора коефіцієнтів b;
stats.p - значення рівня значущості p статистики t Стьюдента для вектора коефіцієнтів b;
stats.resid - вектор залишків;
stats.residp - вектор залишків Пірсона;

stats.residd - вектор залишків узагальненої залишкової суми квадратів;
stats.resida - вектор залишків Анскомбе.

Якщо обчислюється оцінка дисперсії параметра для біноміального закону або розподілу Пуассона, то *stats.s=stats.sfit*. Елементи вектора *stats.se* будуть відрізнятися від їх теоретичних значень на величину *stats.s*.

GLMVAL

Обчислення значень залежної змінної за узагальненою регресійною моделлю

Синтаксис

```
yfit = glmval(b,X,'link')
[yfit,dlo,dhi] = glmval(b,X,'link',stats,clev)
[yfit,dlo,dhi] = glmval(b,X,'link',stats,clev,N,offset,'const')
```

Опис

yfit = glmval(b,X,'link') – функція призначена для обчислення залежної змінної *yfit* для значень незалежної змінної *X* на основі вектора коефіцієнтів лінійної узагальненої регресійної моделі *b* і функції взаємозв'язку *'link'*. У загальному випадку вектор коефіцієнтів регресійної моделі b розраховується за допомогою функції *glmfit*.

Вхідний аргумент *'link'* задає вид взаємозв'язку між параметром розподілу й оцінюваною лінійною комбінацією незалежних змінних $X \cdot b$. Параметр *'link'* задається відповідно до вимог до такого ж вхідного аргументу функції *glmfit*. Передбачено наступні значення *'link'* (табл. 2.3).

Таблиця 2.3

Значення вхідного аргументу *'link'* функції *glmval*

Значеннння *'link'*	Вид залежності
1	2
'identity'	$\mu = xb$
'log'	$\log(\mu) = xb$
'logit'	$\log(\mu/(1-\mu)) = xb$
'probit'	$\text{norminv}(\mu) = xb$
'comploglog'	$\log(-\log(1-\mu)) = xb$
'logloglink'	$\log(-\log(\mu)) = xb$

Закінчення табл. 2.3

1	2
'reciprocal'	$1/\mu = xb$
p (число)	$\mu^p = xb$

Крім наведених вище значень *'link'* можливо задати довільний вид залежності за допомогою *inline* функцій або М-файлів. Правила й приклади використання такого способу – в описі функції *glmfit*.

Вихідний параметр *yfit* є значенням оберненої функції взаємозв'язку лінійної комбінації $X \cdot b$.

[yfit,dlo,dhi] = glmval(b,X,'link',stats,clev) – у цьому варіанті синтаксису функція повертає нижнє *dlo* та верхнє *dhi* відхилення від *yfit* меж довірчого інтервалу параметрів закону розподілу для заданих *b*, *X*, *'link'*. Для розрахунку *dlo*, *dhi* використовується структура *stats*, отримана як вихідний параметр функції *glmfit*. Вхідний параметр *clev* задає довірчу ймовірність для обчислення меж довірчого інтервалу. За замовчуванням *clev* приймається дорівненою 0,95, що відповідає 95 % довірчого інтервалу. Межі довірчого інтервалу обчислюються як *[yfit-dlo, yfit+dhi]*.

[yhat,dlo,dhi]=glmval(beta,X,'link',stats,clev,N,offset,'const') – додаткові вхідні параметри *N, offset, 'const'* потрібні для забезпечення однакових умов з обчисленнями, проведеними за допомогою функції *glmfit.* У використанні біноміального закону в роботі з *glmfit* необхідно задати вхідний параметр *N* відповідному числу повторних незалежних випробувань. Якщо у виклику функції *glmfit* використовувалися параметри *offset* й *'const'*, то необхідно у виклику функції glmval використати ці ж параметри з тими ж значеннями.

CORRCOEF **Обчислення матриці парних коефіцієнтів кореляції**

Синтаксис

```
R = corrcoef(X)
R = corrcoef(x,y)
[R,P]=corrcoef(...)
[R,P,RLO,RUP]=corrcoef(...)
[...]=corrcoef(...,'param1',val1,'param2',val2,...)
```

Опис

R = corrcoef(X) – функція призначена для обчислення матриці парних коефіцієнтів кореляції *R* вибірок, поданих у вигляді матриці *X*. Спостереження розташовуються у рядках матриці *X*, вибірки – у стовпцях.

Обчислення *(i,j)* елемента матриці *R* здійснюється за формулою $R(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)\cdot C(j,j)}}$, де *C = cov(X)* – матриця коваріацій.

R = corrcoef(x,y) – функція призначена для обчислення матриці парних коефіцієнтів кореляції *R* векторів *x* й *y*. Такий результат можна отримати, використавши *corrcoef([x y])*.

[R,P]=corrcoef(...) – функція повертає матриці парних коефіцієнтів кореляції *R* і рівнів значущості *P*, які використовуються у перевірці гіпотези про відсутність кореляції. Кожне значення *P* є значенням імовірності отримати величину коефіцієнта кореляції більшу, ніж обчислене вибіркове значення під дією випадкових факторів, коли істине значення коефіцієнта кореляції дорівнює нулю. Якщо *P(i,j)* менше 0,05, то значення коефіцієнта кореляції *R(i,j)* є значущим.

[R,P,RLO,RUP]=corrcoef(...) – функція повертає матриці парних коефіцієнтів кореляції *R*, рівнів значущості *P*, нижніх *RLO* та верхніх *RUP* меж 95 % довірчих інтервалів коефіцієнтів кореляції.

[...]=corrcoef(...,'param1',val1,'param2',val2,...) – у цьому варіанті синтаксису функції додаткові вхідні параметри визначають таким чином (табл. 2.4).

Таблиця 2.4

Значення додаткових вхідних параметрів функції *corrcoef*

'alpha'	Значення рівня значущості. Довірча ймовірність визначається як *100*(1 - alpha)%*. За замовчуванням рівень значущості дорівнює 0,05, що відповідає 95 % довірчого інтервалу коефіцієнта кореляції
'rows'	Визначає спосіб виключення рядків матриці *X* зі значеннями *NaN* в обчисленні коефіцієнта кореляції. Можливі значення параметра: *'all'* – використовуються всі рядки (значення за замовчуванням), *'complete'* – виключаються рядки зі значеннями *Na*, *'pairwise'* – в обчисленні *R(i,j)* виключаються рядки, що містять *Na* у стовпці *i* або *j*

Значення рівня значущості обчислюються на основі перетворення коефіцієнта кореляції в *t*-статистику з `n-2` ступенями свободи, де *n* – кількість рядків матриці *X*. Межі довірчого інтервалу коефіцієнта кореляції обчислюються на підставі того, що статистика, обчислена як `0,5*log((1+R)/(1-R)),` має асимптотичне наближення до нормального закону з дисперсією, що дорівнює `1/(n-3)`. Обчислені в такий спосіб межі довірчого інтервалу коефіцієнта кореляції є точними з більшими вибірками, коли *X* розподілені за багатомірним нормальним законом. Параметр `'rows'`, який дорівнює `'pairwise'`, може призвести до отримання матриці *R*, яка не буде додатньо визначеною.

Функція `corrcoef` є функцією ядра *MATLAB*.

2.3. Розв'язування типової задачі в середовищі *Matlab*

Задача. За територіями регіонів наводяться дані за 199X р. (табл. 2.5)

Таблиця 2.5

Вихідні дані задачі

Номер регіону	Середньодушовий прожитковий мінімум за день одного працездатного, у. о, *x*	Середньоденна заробітна плата, у. о, *y*
1	78	133
2	82	148
3	87	134
4	79	154
5	89	162
6	106	195
7	67	139
8	88	158
9	73	152
10	87	162
11	76	159
12	115	173

Для початку роботи необхідно створити новий М-файл. Для цього з меню *File* вибрати опцію *New*, а потім M-File.

У вікні, що з'явилося, редактора М-файлів ввести вихідні дані. Це можна зробити двома способами: отримати дані з файлу або ввести дані вручну.

Фрагмент М-файла

```
%% Отримати дані з файла
D = dlmread('lab1.txt');
y = D(:,1)';% в першому стовпці значення залежної змінної
x = D(:,2)';% у другому стовпці - незалежної
%% Ввести дані вручну:
y = [133 148 134 154 162 195 139 158 152 162 159 173];
x = [78 82 87 79 89 106 67 88 73 87 76 115];
```

Використовуючи функцію *GLMFIT* побудувати функцію лінійної парної регресії.

Для висновку функції в робочу область у наочному виді використати функції *subs* і *pretty*.

Фрагмент М-файла

```
%% Визначення рівняння парної лінійної регресії, побудова графіка
[P,dev,stats] = glmfit(x,y);
fprintf('Рівняння лінійної парної регресії:')
y_p=subs(sym('a0 + a1*x'),{'a0','a1'},[P(1) P(2)]);
pretty(y_p)
```

```
>> Рівняння лінійної парної регресії
>> y_p = 0,9204 x + 76,9765
```

Параметр регресії дозволяє зробити висновок, що зі збільшенням середньодушового прожиткового мінімуму на 1 у.о. середньоденна заробітна плата зростає в середньому на 0,92 у. о.

Тісноту лінійного зв'язку оцінюють коефіцієнтом кореляції r_{xy}. Для його обчислення використовують функцію *corrcoef*. Коефіцієнт детермінації обчислюють, використовуючи формулу (2.7) та співвідношення $R^2 = r_{xy}^2$.

Фрагмент М-файла

```
%% коефіцієнт кореляції і коефіцієнт детермінації
Rxy = corrcoef(x,y);
fprintf('Коефіцієнт кореляції r:')
r = Rxy(1,2)
fprintf('Коефіцієнт детермінації R:')
R = r^2
```

```
>> Коефіцієнт кореляції r:
r =

    0,7210

>> Коефіцієнт детермінації R:
R =

    0.5199
```

Оскільки значення коефіцієнта кореляції більше 0,7, то це говорить про наявність сильного лінійного зв'язку між ознаками. Коефіцієнт детермінації дорівнює 0,5199, а це означає, що 51,99 % варіації заробітної плати (*y*) пояснюється варіацією фактору *x* – середньодушового прожиткового мінімуму.

Для обчислення середньої похибки апроксимації вбудованої функції в Matlab 2009*b* немає, однак, використовуючи базові знання основ програмування, можна створити самостійно функцію обчислення $\overline{A}$.

Фрагмент М-файла

```
fprintf('Середня помилка апроксимації:')
n = length(y);% розмір масиву вихідних даних
y_e = stats.resid; % вектор залишків
A = 0;
for i=1:n
    A = A + abs(y_e(i)/y(i));% сума відносин
end;
A = (A/n)*100 % обчислення середньої помилки апроксимації
```

```
>> A =

    5,7521
```

Якість побудованої моделі оцінюється як добра, тому що $\overline{A}$ не перевищує 10 %.

Оцінювання статистичної значущості рівняння регресії в цілому слід провести за допомогою *F*-критерію Фішера з урахуванням присвоєння змінної *df* числа ступенів свободи.

Фрагмент М-файла

```
fprintf('число ступенів свободи')
df=stats.dfe
%% Критерій Фішера
fprintf('Критерій Фішера')
F_p = r^2/(1-r^2)*df
```

```
>> число ступенів свободи

df =

    10

>> Критерій Фішера

F_p =

   10.8280
```

Табличне значення критерію з рівнем значущості $\alpha = 0{,}05$ і ступенях свободи $m_1 = 1$ і $m_2 = 10$ становить $F_{табл} = 4{,}96$. Оскільки $F_{розр} > F_{табл}$, то рівняння регресії є статистично значущим.

Оцінювання статистичної значущості параметрів регресії та кореляції провести за допомогою *t*-статистики Стьюдента.

Фрагмент М-файла

```
fprintf('значення статистики t Стьюдента для параметрів рівняння регресії')
t_a0 = stats.t(1)
t_a1 = stats.t(2)
% Перевірка rxy за критерієм Стьюдента:
t_rxy = r/sqrt((1-r^2)/df)
```

```
>> значення статистики t Стьюдента для параметрів рівняння регресії

t_a0 =

    3,1793
t_a1 =
```

```
    3,2906

t_rxy =

    3,2906
```

Фактичні значення *t-статистики* перевершують табличне значення:

$$t_{a_0\ розр} = 3{,}17 > t_{табл} = 2{,}3; \quad (2.24)$$

$$t_{a_1\ розр} = 3{,}29 > t_{табл} = 2{,}3; \quad (2.25)$$

$$t_{r_{xy}\ розр} = 3{,}29 > t_{табл} = 2{,}3, \quad (2.26)$$

тому параметри a_0, a_1 і r_{xy} не випадково відрізняються від нуля, а статистично значущі.

Отримані оцінки рівняння регресії дозволяють використати його для прогнозу. Якщо прогнозне значення прожиткового мінімуму складе: $x_{прогн} = \frac{x \cdot 107}{100\,\%} = 91{,}6$ у. о., тоді індивідуальне прогнозне значення заробітної плати легко обчислити за допомогою наступного програмного коду в Matlab.

Фрагмент М-файла

```
%% Прогнозування
fprintf('Виконання прогнозу на підставі лінії регресії:')
xnew = 91.6;
[yfit,y1,y2] = glmval(P,xnew,'identity',stats,0.95);
ynew = yfit
```

```
>>Виконання прогнозу на основі рівняння регресії:
ynew =

  161,2879
```

Таким чином, з прогнозним значенням прожиткового мінімуму $x_{прогн} = 91{,}6$ у. о. індивідуальне прогнозне значення заробітної плати складе: $y_{прогн} = 161{,}3$ у. о.

Необхідно знайти довірчі інтервали для коефіцієнтів регресії та середнього значення змінної *y* за заданим значенням *x*. Довірча ймовірність *p* = 95 %. Для розв'язання цієї задачі використовують функцію *GLMVAL*.

Фрагмент М-файла

```
fprintf('Інтервали прогнозу для коефіцієнтів регресії:')
interval_a0=subs(sym('a01 < a0 < a02'),{'a01','a02'},[P(1)-
t_tabl*S_a0 P(1)+t_tabl*S_a0])
interval_a1=subs(sym('a11 < a1 < a12'),{'a11','a12'},[P(2)-
t_tabl*S_a1 P(2)+t_tabl*S_a1])
```

```
>>Інтервали прогнозу для коефіцієнтів регресії:

interval_a0 =

21,2899 < a0 < 132,6631

interval_a1 =

0,2771 < a1 < 1,5638
```

Фрагмент М-файла

```
fprintf('Інтервали прогнозу для y:')
yl = yfit-y1;
yt = yfit+y2;
interval_y1=subs(sym('y1 < y < y2'),{'y1','y2'},[yl yt])
fprintf('Інтервал прогнозу для індивідуального значення y:')
S = sqrt(sum(e.^2)/df);
S11 = sqrt(1 + 1/n + (xnew-sum(x)/n)^2 / (sum(x.^2) - (sum(x))^2 /
n));
interval_y3=subs(sym('y1 < y < y2'),{'y1','y2'},[ynew-t_tabl*S*S11
ynew+t_tabl*S*S11])
```

```
>> Інтервали прогнозу для y:

interval_y1 =

152,3874 < y < 170,1884

>> Інтервал прогнозу для індивідуального значення y:

interval_y3 =

130.9969 < y < 191.5789
```

Аналіз верхньої та нижньої меж довірчих інтервалів призводить до висновку про те, що з імовірністю $p = 0{,}95$ параметри a_0 й a_1, перебуваючи в зазначених межах, не набувають нульових значень, тобто є статистично значущими й істотно відмінні від нуля.

Довірчий інтервал для середнього значення змінної *в* з заданим значенням x = 91,6 у. о. дорівнює (152,4; 170,2). Довірчий інтервал для індивідуального значення змінної *в* з заданим значенням x = 91,6 у. о. дорівнює (130,9; 191,6).

Помітно, що довірчий інтервал для індивідуальних значень змінної *в* ширший довірчого інтервалу для середнього значення змінної *y*.

Таким чином, виконаний прогноз середньомісячної заробітної плати є надійним і перебуває в межах від 130,9 у. о. до 191,6 у. о. (рис. 2.1).

Для закінчення розв'язання задачі треба побудувати на одному графіку вихідні дані та теоретичну пряму, використовуючи функцію *plot*.

Фрагмент М-файла

```
%% Графік експериментальних даних і лінія регресії
y_p = P(1) + P(2)*x;
plot(x,y,'mo',x,y_p,'k')
title('Експериментальні дані і лінія регресії')
xlabel('x'); ylabel('y');
text(77, 160, '\leftarrow Експериментальні дані')
text(100, 169, '\leftarrow Лінія регресії')
grid on
```

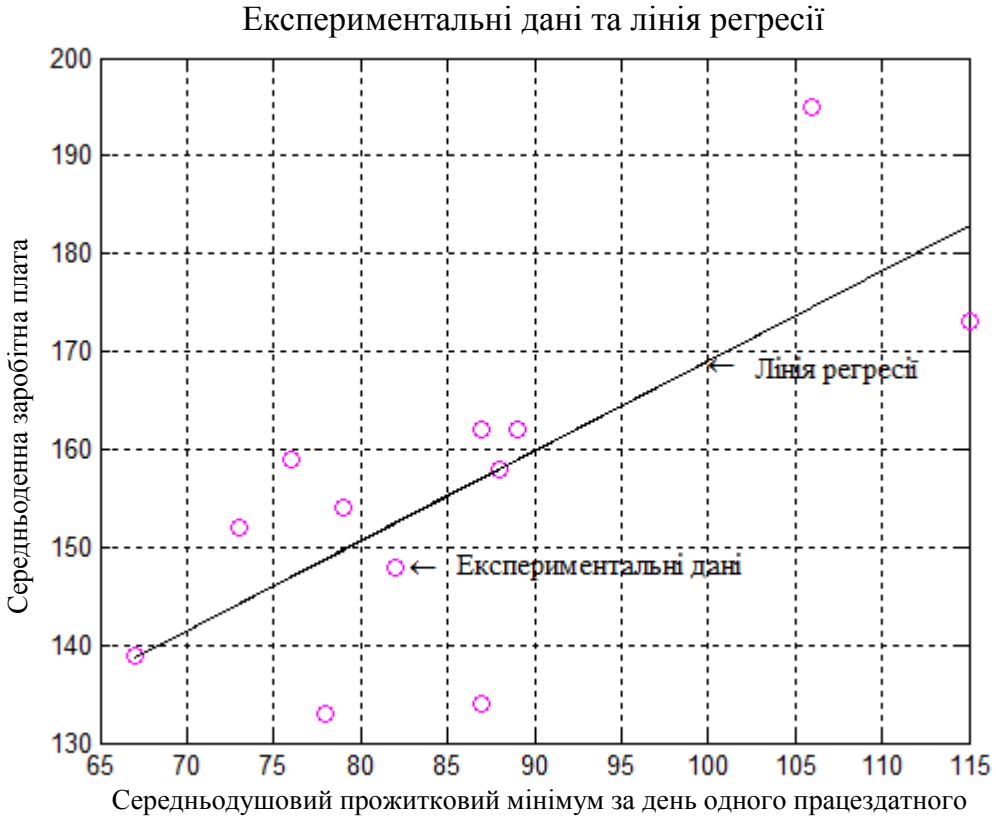


Рис. 2.1. **Графічне відображення експериментальних даних та лінії регресії**

Запитання для самоперевірки

1. У чому полягає кореляційно-регресійний аналіз?

2. Що означають коефіцієнти в парній лінійній регресії?

3. За якою формулою обчислюється лінійний коефіцієнт парної кореляції r_{xy}?

4. Як побудувати довірчий інтервал для лінійного коефіцієнта парної кореляції?

5. Який критерій використовують для оцінювання значущості коефіцієнта кореляції?

6. Як обчислюється та що показує коефіцієнт детермінації?

7. Як перевіряється значущість рівняння регресії?

8. Що означає середня помилка апроксимації?

9. У чому відмінність точкового прогнозу від інтервального?

10. Як побудувати довірчі інтервали для параметрів рівняння регресії?

11. Як побудувати довірчий інтервал для середнього й індивідуального значень змінної y?

Завдання до лабораторної роботи

За територіями регіону наведені дані за 199X р. (p_1 – кількість букв у повному імені, p_2 – кількість букв у прізвищі) (табл. 2.6).

Таблиця 2.6

Вихідні дані задачі

Номер регіону	**Середньодушовий прожитковий мінімум у день одного працездатного, у. о, x**	**Середньоденна заробітна плата, у. о, y**
1	$78 + p_1$	$133 + p_2$
2	$82 + p_2$	$148 - p_1$
3	$87 + p_1$	$134 - p_2$
4	$79 + p_1$	$154 + p_2$
5	$89 + p_2$	$162 - p_1$
6	$106 - p_2$	$195 - p_1$
7	$67 + p_2$	$139 + p_1$
8	$88 + p_1$	$158 - p_1$
9	$73 + p_2$	$152 - p_1$
10	$87 + p_2$	$162 + p_1$
11	$76 - p_1$	$159 - p_2$
12	$115 + p_2$	$173 - p_1$

Необхідно оформити письмовий звіт, що містить розв'язання наступних задач:

1. Побудувати лінійне рівняння парної регресії *y* за *x* .

2. Обчислити лінійний коефіцієнт парної кореляції, коефіцієнт детермінації та середню помилку апроксимації.

3. Оцінити статистичну значущість рівняння регресії в цілому й окремих параметрів регресії та кореляції за допомогою *F*-критерію Фішера та *t*-критерію Стьюдента.

4. Виконати прогноз заробітної плати *y* з прогнозним значенням середньодушового прожиткового мінімуму *x*, що становить 107 % від середнього рівня.

5. Обчислити довірчі інтервали для параметрів регресії, виконати точковий та інтервальний прогноз за рівнянням лінійної регресії.

6. Побудувати графік вихідних даних і рівняння регресії.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

Лабораторна робота 3
Нелінійні моделі й їх лінеаризація

Мета роботи: набути компетентності у побудові нелінійної моделі за допомогою вбудованих функцій Matlab (*Curve Fitting Toolbox*).

Основні задачі лабораторної роботи:

1. Знайти рівняння регресії, використовуючи вбудовані параметричні моделі.

2. Знайти рівняння регресії, використовуючи вбудовані непараметричні моделі: інтерполяційні та згладжувальні сплайни.

3. Провести дисперсійний аналіз.

4. Порівняти отримані моделі за допомогою коефіцієнта детермінації.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

3.1. Основні поняття нелінійної регресії

Багато залежностей в економіці не є лінійними, і тому їх моделювання лінійними рівняннями недоцільне. Наприклад, для дослідження залежності попиту *y* на деякий товар від ціни *x* даного товару в певних випадках можна обмежитись лінійним рівнянням регресії. Проте якщо необхідно визначити

еластичність попиту залежно від ціни, то необхідно обчислити логарифмічну модель. В аналізі залежності витрат y від обсягу випуску x найбільше обґрунтованою є логарифмічна модель. У розгляді виробничих функцій лінійна модель не є реальною, в даному випадку використовують степеневі моделі. Наприклад, широко відома виробнича функція Кобба − Дугласа $y = AK^{\alpha}B^{\beta}$, де y − обсяг випуску, K та B − витрати капітала і праці, відповідно, A, α, β − параметри моделі. Достатньо широко використовують в сучасному економічному аналізі й інші моделі, як, наприклад, обернена й експоненціальна моделі.

Регресії, нелінійні за пояснювальною змінною, можуть мати різний вигляд:

1) поліноми різних ступенів, наприклад: $y = a + b_1 x + b_2 x^2 + b_3 x^3 + \varepsilon$;

2) рівнобічна гіпербола $y = a + \frac{b}{x} + \varepsilon$.

В економічних задачах часто використовують регресії, нелінійні за параметрами, що оцінюються, такого вигляду:

1) степенева $y = a \cdot x^b \cdot \varepsilon$;

2) показникова $y = a \cdot b^x \cdot \varepsilon$;

3) експоненціальна $y = e^{a+b\cdot x} \cdot \varepsilon$.

Відомо, що для успішного використання МНК бажано, щоб модель була лінійною відносно параметрів.

Для оцінювання параметрів нелінійних моделей використовують два підходи.

Перший підхід заснований на *лінеаризації* моделі та полягає в тому, що за допомогою підходящих перетворень вихідних змінних досліджувану залежність подають у вигляді лінійного співвідношення між перетвореними змінними.

В табл. 3.1 наведена лінеаризація різних типів рівнянь регресії.

Таблиця 3.1

Лінеаризація найбільш використовуваних типів рівнянь регресії

Вид функції	**Лінеаризація**	**Параметри рівняння регресії**	**Шукане рівняння**
1	2	3	4
Степенева $y = a_0 x^{a_1}$	$X = \ln x$, $Y = \ln y$, $A_0 = \ln a_0$, $A_1 = a_1$.	$A_1 = \frac{\overline{XY} - \overline{Y} \cdot \overline{X}}{\overline{X^2} - \overline{X}^2}$ $A_0 = \overline{Y} - A_1 \cdot \overline{X}$	$Y = A_0 + A_1 X$ Перехід до початкових змінних здійснюється в такий спосіб: $\tilde{y} = e^Y$, $x = e^X$, $a_0 = e^{A_0}$, $a_1 = A_1$.

Закінчення табл. 3.1

1	2	3	4
Показникова $y = a_0 a_1^{\,x}$	$X = x,$ $Y = \ln y,$ $A_0 = \ln a_0,$ $A_1 = \ln a_1.$	$A_1 = \dfrac{\overline{xY} - \bar{Y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}$ $A_0 = \bar{Y} - A_1 \cdot \bar{x}$	$Y = A_0 + A_1 X$ Перехід до початкових змінних здійснюється в такий спосіб: $\tilde{y} = e^Y,\ \ x = X,$ $a_0 = e^{A_0},\ a_1 = e^{A_1}.$
Обернена $y = \dfrac{1}{a_0 + a_1 x}$	$X = x,$ $Y = 1/y,$ $A_0 = a_0,$ $A_1 = a_1.$	$A_1 = \dfrac{\overline{xY} - \bar{Y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}$ $A_0 = \bar{Y} - A_1 \cdot \bar{x}$	$Y = A_0 + A_1 X$ Перехід до початкових змінних здійснюється в такий спосіб: $\tilde{y} = 1/Y,\ \ x = X,$ $a_0 = A_0,\ a_1 = A_1.$
Напівлогарифмічна $y = a_0 + a_1 \ln x$	$X = \ln x,$ $Y = y,$ $A_0 = a_0,$ $A_1 = a_1.$	$A_1 = \dfrac{\overline{Xy} - \bar{y} \cdot \bar{X}}{\overline{X^2} - \bar{X}^2}$ $A_0 = \bar{y} - A_1 \cdot \bar{X}$	$Y = A_0 + A_1 X$ Перехід до початкових змінних здійснюється в такий спосіб: $\tilde{y} = Y,\ \ x = e^X,$ $a_0 = A_0,\ a_1 = A_1.$
Гіперболічна $y = a_0 + \dfrac{a_1}{x}$	$X = 1/x,$ $Y = y,$ $A_0 = a_0,$ $A_1 = a_1.$	$A_1 = \dfrac{\overline{Xy} - \bar{y} \cdot \bar{X}}{\overline{X^2} - \bar{X}^2}$ $A_0 = \bar{y} - A_1 \cdot \bar{X}$	$Y = A_0 + A_1 X$ Перехід до початкових змінних здійснюється в такий спосіб: $\tilde{y} = Y,\ \ x = 1/X,$ $a_0 = A_0,\ a_1 = A_1.$
Експоненціальна $y = a_0 e^{a_1 \cdot x}$	$X = x,$ $Y = \ln y,$ $A_0 = \ln a_0,$ $A_1 = a_1.$	$A_1 = \dfrac{\overline{xY} - \bar{Y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}$ $A_0 = \bar{Y} - A_1 \cdot \bar{x}$	$Y = A_0 + A_1 X$ Перехід до початкових змінних здійснюється в такий спосіб: $\tilde{y} = e^Y,\ \ x = X,$ $a_0 = e^{A_0},\ a_1 = A_1.$

Другий підхід звичайно застосовують у випадках, коли підібрати відповідне лінеаризаційне перетворення не вдається. Тоді використовують *методи нелінійної оптимізації* на основі вихідних змінних.

Побудоване рівняння нелінійної регресії доповнюється показником кореляції, а саме − *індексом кореляції*:

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_i - \tilde{y}_{xi})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}},\ 0 \le \rho_{xy} \le 1. \qquad (3.1)$$

Чим ближча величина даного показника, тим тісніший зв'язок між розглянутими ознаками та надійніше побудоване рівняння регресії.

Оцінку якості побудованої моделі дає індекс детермінації ρ_{xy}^2.

Коефіцієнт детермінації

$$R^2 = \rho_{xy}^2 \tag{3.2}$$

– квадрат індекса кореляції – характеризує частку дисперсії, що пояснює регресією, у загальній дисперсії результівної ознаки *y*.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (x_i - y_i)}{\sum_{i=1}^{n} (y_i - \bar{y})} \tag{3.3}$$

Чим ближчий коефіцієнт детермінації до 1, тим вища якість рівняння регресії та повніше воно пояснює поведінку результівної ознаки.

3.2. Теоретичні відомості про функції *Matlab*, які використовуються в даній лабораторній роботі

Curve Fitting Toolbox – це пакет розширення *MATLAB* для різних прикладних задач налагоджування, апроксимації й інтерполяції даних. Містить у собі інтерактивні засоби попередньої обробки даних, порівняння стандартних моделей та розроблення моделей користувача, налагоджування за допомогою стандартних і робастних методів й аналізу якості апроксимації.

До складу *Curve Fitting Toolbox* входить додаток ***cftool*** із графічним інтерфейсом користувача, що дозволяє здійснювати всі вищеперераховані дії та функції, призначені для визначення параметричної моделі, підбора параметрів, аналізу придатності наближення, операцій з ним і графічним відображенням результату.

Імпорт даних у *MATLAB*. Для прочитування даних з файла в *MATLAB* існує кілька способів, одним із яких є використання вбудованих функцій *MATLAB*.

Наприклад, дані записані в текстовому файлі у два стовпчики і як роздільник десяткових розрядів використовується крапка. Для запису стовпців файла у векторі *MATLAB* можна скористатися функцією `dlmread`.

DLMREAD **Читання ASCII-файлів**

Синтаксис

```
RESULT = DLMREAD(FILENAME)

RESULT = DLMREAD(FILENAME,DELIMITER)

RESULT = DLMREAD(FILENAME,DELIMITER,R,C)

RESULT = DLMREAD(FILENAME,DELIMITER,RANGE)
```

Опис

Dlmread прочитує файл *ASCII* під ім'ям *'FILENAME'*.

RESULT = DLMREAD(FILENAME,DELIMITER) – прочитує дані з файла *FILENAME* з *ASCII*-роздільником, використовуючи роздільник *DELIMITER*, у масив *RESULT*. Використайте *'\t'*, щоб визначити символ табуляції як роздільник.

RESULT = DLMREAD(FILENAME,DELIMITER,R,C) – прочитує дані з файла *FILENAME* з *ASCII*-роздільником, використовуючи роздільник *DELIMITER*, у масив *RESULT*, починаючи зі зміщення *R* (у рядках) і *C* (у стовпцях). Параметри *R* і *C* відраховують починаючи з нуля, так що *R = 0*, *C = 0* відповідає першому значенню у файлі.

RESULT = DLMREAD(FILENAME, DELIMITER, RANGE) – імпортує індексований або іменований діапазон даних з роздільниками у форматі ASCII. Для використання діапазону осередків потрібно визначити параметр *RANGE* у вигляді *range = [Верхній Рядок, Лівий Стовпець, Нижній Рядок, Правий Стовпець]*.

Для подальшої роботи в додатку *cftool* варто створити два вектори *x* й *y*, що містять, відповідно, перший і другий стовпці масиву *RESULT*. Для цього варто ввести в М-файлі дві команди, у яких створюється індексація: для обігу до всіх елементів першого та другого стовпців масиву *RESULT* використана двокрапка (індексація двокрапкою), а номер стовпця, що виділяється у вектор, задається числом.

XLSREAD **Читання даних з *Excel***

Синтаксис

```
[NUMERIC,TXT,RAW]=XLSREAD(FILE)
```

Опис

Для прочитування даних з *Excel* в *MATLAB* є спеціальна функція `xlsread`, що допускає кілька способів виклику доцільно розглянути основні з них.

Указівка імені книги *Excel* в апострофах як вхідний аргумент функції `xlsread` призводить до прочитування даних з першого аркуша книги. Якщо перший аркуш не містить даних, то функція `xlsread` поверне порожній масив. У прочитуванні даних з аркуша книги *Excel* ігноруються верхні рядки та праві стовпці, що містять текст. Числові дані, отримані в ланцюгах аркуша в результаті обчислення за формулами, прочитуються функцією `xlsread` як звичайні числові константи, введені в ланцюги.

Якщо, наприклад, книга *ExpData.xls* містить три аркуші, які називаються *Experiment*1, *Experiment*2, *Experiment*3, то команда `A=xlsread('ExpData.xls')` призведе до появи в робочому середовищі масиву `A`.

Якщо діапазон значень на першому аркуші містить текст і порожні ланцюги, то замість них у масиві, що повертає функцією `xlsread`, буде перебувати `NaN`.

Для прочитування даних із книги *Excel* не тільки з першого, а й з довільного аркуша, досить указати номер аркуша або його ім'я (в апострофах) у якості другого вхідного аргументу:

```
A = xlsread('ExpData.xls',2)
```

або

```
A = xlsread('ExpData.xls','Experiment2').
```

Функція `xlsread` дозволяє вказати діапазон даних, які потрібно прочитати. Якщо діапазон даних зазначений у якості її другого вхідного аргументу, то відбувається прочитування даних із цього діапазону на першому аркуші книги Ecxel. Якщо в другому вхідному аргументі `xlsread` указати ім'я аркуша, а діапазон даних – у третьому, то буде виконуватись прочитування

даних із заданого діапазону цього аркуша. Якщо, наприклад, дані на аркуші *Experiment2* розташовані в діапазоні B2:C4, то для їх прочитування варто застосувати:

```
A = xlsread('ExpData.xls','Experiment2','B2:C4')
```

Функція *xlsread* надає зручну можливість інтерактивного виділення діапазону прочитуваних даних. Якщо в якості її другого вхідного аргументу вказати -1, то в *Excel* відкриється книга, ім'я якої задане в першому вхідному аргументі функції *xlsread,* і з'явиться діалогове вікно. Залишилося вибрати аркуш, на аркуші виділити діапазон даних і натиснути кнопку OK у цьому вікні. Дані з обраного діапазону повернуться у вихідному аргументі функції *xlsread*.

Якщо потрібно вважати дані із всіх аркушів книги *Excel* у робоче середовище *MATLAB*, то зручно занести їх у структуру, поля якої будуть збігатися з назвами аркушів книги, а їх зміст буде масивами числових даних, розташованих на відповідних аркушах книги.

Вікно додатка *cftool*. Для запуску додатка *cftool* достатньо набрати в М-файлі його ім'я *cftool*:

```
>> cftool
```

З'являється вікно додатка (на рис. 3.1 наведена тільки частина вікна *cftool* із вказівкою призначення основних його компонентів).

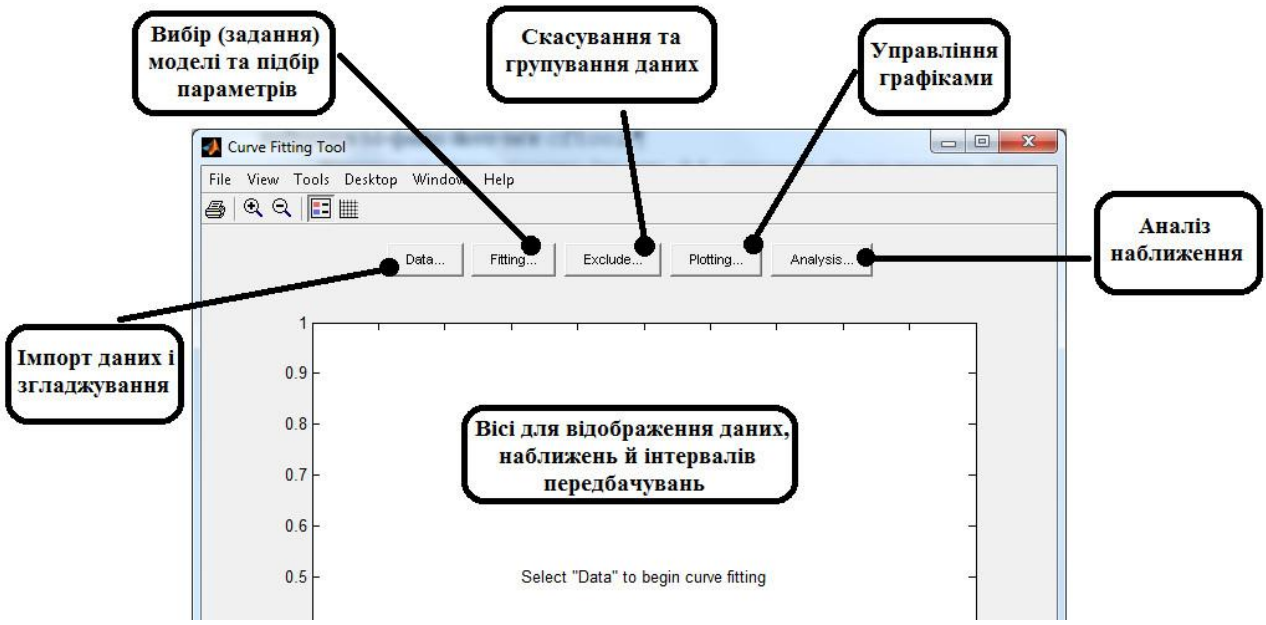


Рис. 3.1. **Вікно додатка *cftool***

Основні етапи розв'язування задачі про підбір параметрів параметричної моделі, що наближає дані, у додатку `cftool` такі:

1. Імпорт даних (кнопка ***Data***) (рис. 3.2).

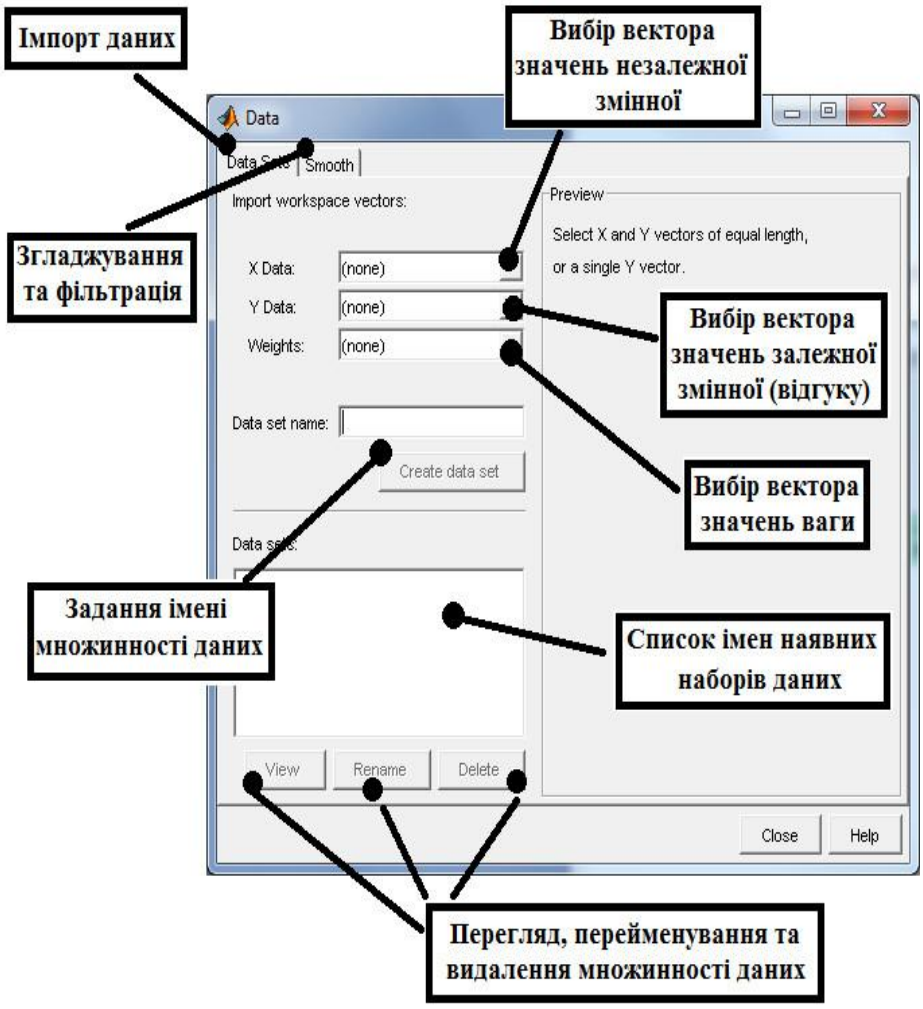


Рис. 3.2. **Діалогове вікно *Data***

2. Побудова, за необхідності, правил виключення деяких значень, виключення вручну в таблиці або угруповання даних для наближення їх частин різними моделями (кнопка ***Exclude***).

3. Вибір стандартної параметричної або непараметричної моделі, що входить в *Curve Fitting Toolbox*, або створення власної моделі, підбір параметрів з попередньою вказівкою їх меж і початкових наближень, цільової функції та методів розв'язання, а також перегляд отриманих значень й інформації про придатність отриманого наближення (кнопка ***Fitting***) (рис. 3.3).

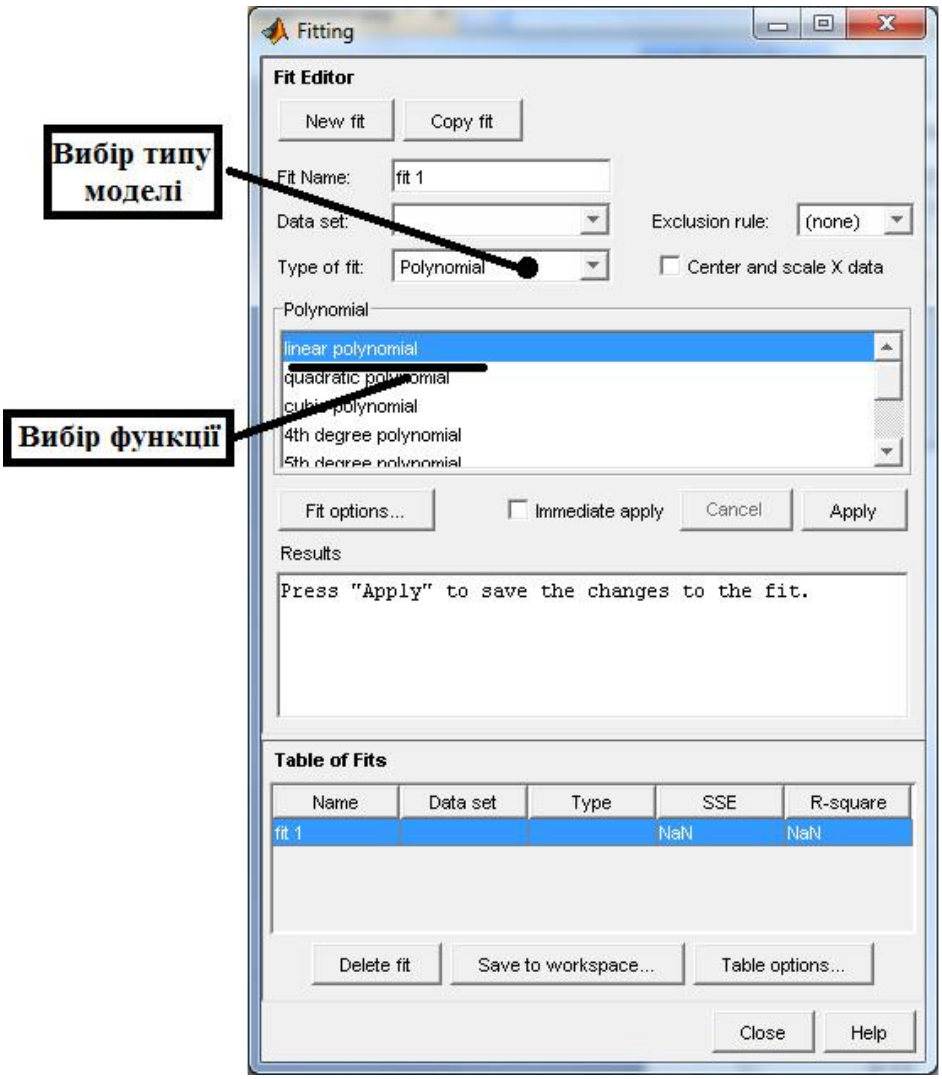


Рис. 3.3. **Діалогове вікно *Fitting***

4. Аналіз даних, що включає обчислення отриманого наближення в заданих точках (включаючи екстраполяцію), його інтегрування та диференціювання (кнопка ***Analysis***).

Крім того, можливо:

1. Залишити тільки ті графіки даних і моделей, які потрібні в цей момент (кнопка ***Plotting***).

2. Форматувати графіки даних і побудованих параметричних моделей (контекстне меню ліній графіків, інструменти вікна ***cftool***, меню ***Tool***).

3. Відобразити графічно інтервали прогнозованих значень із заданою ймовірністю (меню ***View***, пункти ***Prediction Bounds, Confidence Level***).

4. Відобразити графічно залишки (меню ***View***, пункт ***Residuals***).

5. Експортувати наближення та результати їх аналізу в робоче середовище MATLAB (ця можливість є у вікнах, у яких будується наближення та проводиться аналіз).

6. Проводити згладжування та фільтрацію даних (кнопка ***Data***). Однак, необхідно мати на увазі, що згладжування скасовує стандартне припущення регресійного аналізу про те, що розподіл похибки у вихідних даних підпорядковується нормальному закону. Якщо побудовано досить гарну модель, то залишки (різниця значень даних і наближення) також повинні підпорядковуватись нормальному закону. Тому згладжування варто використати як інструмент для отримання первісного припущення про можливу параметричну модель у випадку зашумлених даних, а будувати модель треба для незгладжених вихідних даних.

7. Згенерувати файл-функцію, яку можна використати згодом автономно від додатка ***cftool*** для отримання побудованого в додатку ***cftool*** наближення (меню ***File***, пункт ***Generate M-file***).

8. Зберегти процес і під час наступних запусків додатка *cftool* відновити її (меню ***File***, пункти ***Save Session, Load Session***), а також видалити всі дані й отримані результати (меню ***File***, пункт ***Clear Session***).

9. Вивести результати в окреме графічне вікно (меню ***File***, пункт ***Print to Figure***).

10. Надрукувати результати (меню ***File***, пункт ***Print***).

Експорт результатів у робоче середовище. Додаток ***cftool*** дозволяє експортувати отримані результати в робоче середовище MATLAB, причому можна експортувати згладжені дані, саму параметричну модель і критерії її придатності, інформацію про хід алгоритму пошуку значень параметрів і результати деяких операцій з отриманою параметричною моделлю.

Експорт згладжених даних. Після згладжування дані можуть бути експортовані в робоче середовище MATLAB. Це робиться за допомогою кнопки ***Save to workspace*** вкладки ***Smooth*** діалогового вікна ***Data***. У вікні, що з'являється, варто вказати ім'я структури (глобального змінного робочого середовища), у якій будуть записані згладжені дані, наприклад *smdata*. Тоді поля *x* й *y* структури *smdata* будуть містити:

```
smdata.x - точки, у яких задані вихідні дані;
smdata.y - згладжені дані.
```

Експорт параметричної моделі. Після підбора параметрів у параметричній моделі результати можуть бути експортовані в робоче середовище

MATLAB. Експорт здійснюється за допомогою кнопки *Save to workspace* діалогового вікна *Fitting*. Пропонується зберегти:

саму параметричну модель (відмітка *Save fit to MATLAB object named*);

значення критеріїв придатності наближення (відмітка *Save goodness of fit to MATLAB struct named*);

інформацію, що повертає оптимізаціонним алгоритмом (відмітка Save fit output to MATLAB struct named).

Параметрична модель є об'єктом (див. Функції *Curve Fitting Toolbox*). Інформація про неї містить:

тип моделі й її вид;

знайдені значення параметрів разом з довірчими інтервалами;

sse – сума квадратів похибок (*sum of squares due to error*);

rsquare – критерій *R*-квадрат, квадрат змішаної кореляції (*coefficient of determination*);

dfe – число ступенів свободи, тобто різниця між числом даних і параметрів моделі (*degrees of freedom*);

adjrsquare – уточнений критерій *R*-квадрат (*degree-of-freedom adjusted coefficient of determination*);

rmse – корінь із середнього для квадрата похибки (*Root mean Squared Error*).

Інформація, що повертається оптимізаціонним алгоритмом, зберігається в структурі з полями:

- numobs - число даних;
- numparam - число параметрів у параметричній моделі;
- residuals - вектор незв'язань;
- Jacobian - матриця Якобі, використовувана при мінімізації суми квадратів незв'язок;
- exitflag - відмітка, що містить інформацію про причину зупинки обчислювального алгоритму підбора параметрів моделі:

якщо exitflag більше нуля, то розвязок задачі мінімізації суми квадратів незв'язань успішно знайдено;

якщо exitflag дорівнює нулю, то було досягнуто максимально допустиме число викликів мінімізаційної функції;

якщо exitflag менше нуля, то була виявлена розбіжність ітераційного процесу;

iterations - число ітерацій, зроблених під час мінімізації.

funcCount - число викликів мінімізаційної функції;

algorithm - використаний у мінімізації алгоритм, або QR-розкладання для лінійної параметричної моделі.

Експорт результатів аналізу наближення. Результати аналізу побудованого в додатку *cftool* наближення даних параметричною моделлю (довірчі смуги для наближення та даних, значення моделі в заданих точках, значення її першої та другої похідних і значення інтеграла зі змінною верхньою межею) можуть бути експортовані в робоче середовище за допомогою кнопки *Save to workspace* діалогового вікна *Analysis*.

Результати зберігаються в структурі з полями:

xi - абсциси точок, у яких виконувався аналіз наближення даних параметричною моделлю;

boundtype - тип довірчої смуги (для даних або для наближення параметричною моделлю) ;

lower - ординати нижньої межі довірчої смуги для даних або для наближення в точках *xi*;

upper - ординати верхньої межі довірчої смуги для даних або для наближення в точках *xi*;

conflevel - рівень імовірності, якому відповідає довірча смуга для даних або для наближення;

yfit - значення параметричної моделі в заданих точках *xi*;

d2ydx2 - значення другої похідної від параметричної моделі в заданих точках *xi*;

dydx - значення першої похідної від параметричної моделі в заданих точках *xi*;

integral - значення інтеграла зі змінною верхньою межею від параметричної моделі в заданих точках *xi*;

integralstart - нижня межа інтегрування.

3.3. Розв'язування типової задачі в середовищі Matlab

Задача. На підставі даних обсягу продажів торгового дому за 12 місяців необхідно побудувати регресійну модель залежності обсягів продажів від часу (табл. 3.2).

Таблиця 3.2

Дані обсягу продажів торгового дому, у. о.

Місяці	***t***	y_t
Січень	1	200
Лютий	2	310
Березень	3	320
Квітень	4	260
Травень	5	190
Червень	6	210
Липень	7	310
Серпень	8	410
Вересень	9	430
Жовтень	10	370
Листопад	11	300
Грудень	12	320

Для початку роботи необхідно створити новий М-файл. Для цього з меню *File* вибрати опцію *New*, а потім M-File.

У вікні, що з'явилося, редактора М-файлів ввести вихідні дані. Це можна зробити двома способами: отримати дані з файлу або ввести дані вручну.

Фрагмент М-файла

```
%% Нелінійна регресія
clc % Очистити командне вікно
clear all
%% Отримати дані з файла
data = dlmread('lab2.txt');
x = data(:,1);% у першому стовпці значення незалежної змінної
y = data(:,2);% у другому стовпці - залежній
%% Ввести дані вручну:
x = 1:1:12;
y = [200 310 320 260 190 210 310 410 430 370 300 320];
%% Побудуємо ці точки на графіку
plot(x, y, 'o')
plot(x, y, 'o')
title('Експериментальні дані')
xlabel('x'); ylabel('y');
grid on
```

У результаті роботи М-файла будуть побудовані всі експериментальні дані, що знаходились у масивах *x* і y (рис. 3.4).

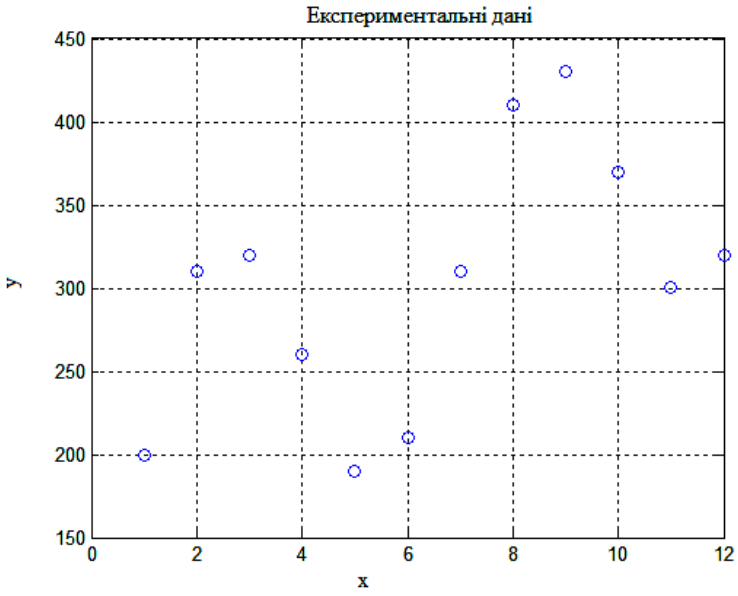


Рис. 3.4. **Вхідні дані задачі**

Для запуску додатка *cftool* набрати в М-файлі його ім'я *cftool* (рис. 3.6).

Фрагмент М-файла

```
%% Curve Fitting Toolbox викликається за допомогою команди
cftool
```

У діалоговому вікні *Curve Fitting Tool*, що з'явилося, натиснути кнопку *Data* для імпортування даних. У цьому діалоговому вікні *Data* заповнити відповідні поля (рис. 3.5).

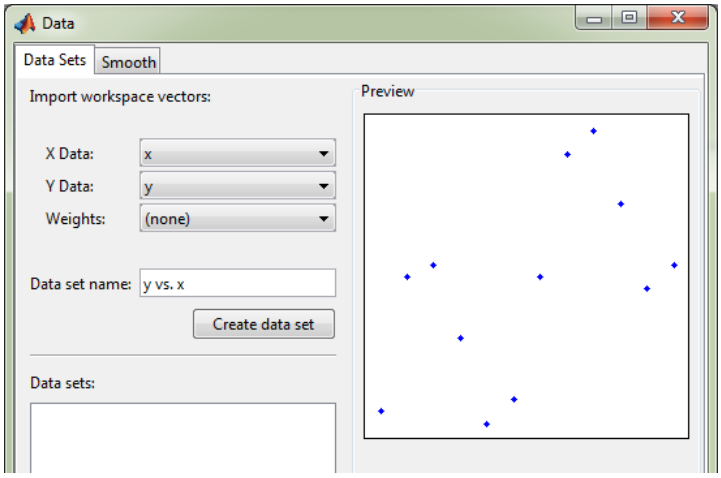


Рис. 3.5. **Додаток *sftool* з вихідними даними**

Після вибору векторів з даними варто задати ім'я цього масива даних. Для цього треба ввести його ім'я в рядок ***Data set name*** та натиснути кнопку ***Create data set*** (вона стає доступною після вибору векторів, що містять дані).

Зі створеною множиною даних можна проробити наступні операції (попередньо слід виділити його ім'я в списку ***Data sets***):

1. Відобразити таблицю даних разом із графіком в окремому вікні, для чого нажати кнопку ***View*** (рис. 3.6).

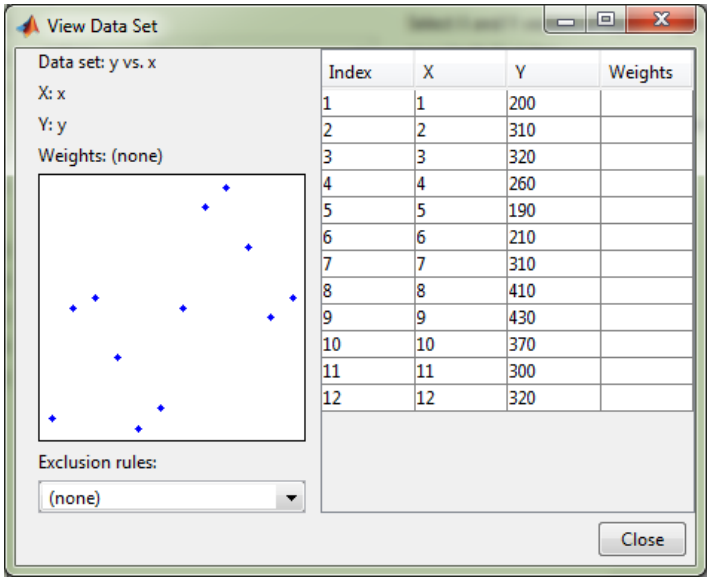


Рис. 3.6. **Діалогове вікно *Data* з вихідними даними**

2. У списку ***Exclusion rules*** вікна ***View Data Set*** можна вибирати правила виключення, поки він порожній, оскільки ніяких правил не задано.

3. Перейменувати виділений набір даних, натиснувши кнопку ***Rename***, після чого з'явиться діалогове вікно, у яке потрібно буде ввести нове ім'я (поки цього робити не потрібно, продовжити роботу з ним).

4. Видалити виділений набір даних, натиснувши кнопку ***Delete*** (цього, мабуть, теж робити зараз не потрібно).

Множина даних створена, його графік відобразився на осях основного вікна додатка *cftool*. У вікні ***Data*** можна також здійснювати згладжування даних (вкладка ***Smooth***), вибираючи різні способи згладжування.

У *Curve Fitting Toolbox* реалізовано три способи згладжування:

ковзне середнє (*Moving average filtering*);

зважена локальна регресія;

фільтр Савицького – Голея.

У даній лабораторній роботі згладжування не розглядається.

Для переходу до діалогового вікна, призначеного для вибору моделі й підбора параметрів, варто нажати кнопку ***Fitting*** в основному вікні додатка ***cftool***. З'являється діалогове вікно ***Fitting***, у якому варто нажати кнопку ***New fit,*** після чого всі елементи керування даного вікна стають доступними.

I. Параметричні моделі

1. Поліноміальні моделі (Polynomials) (рис. 3.7).

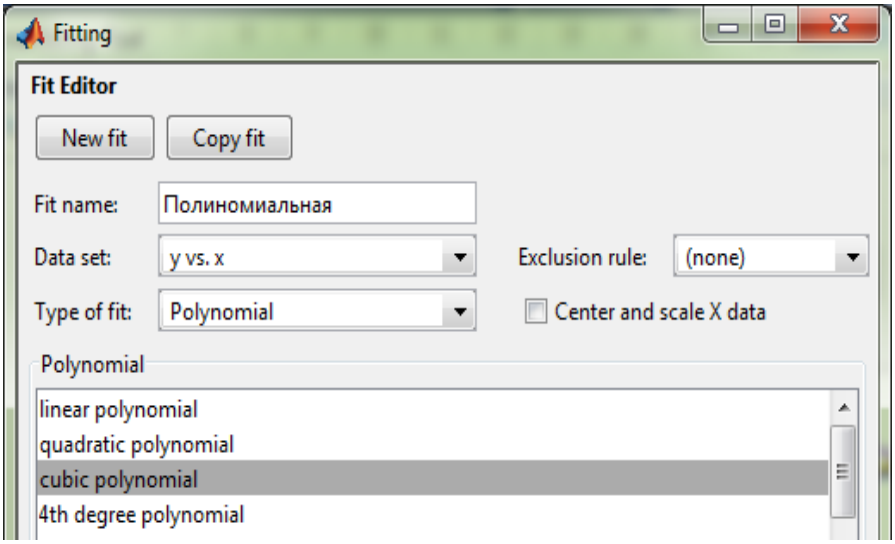


Рис. 3.7. **Вибір поліноміальної моделі діалогового вікна *Fitting***

```
Linear model Poly3:
     f(x) = p1*x^3 + p2*x^2 + p3*x + p4
Coefficients (with 95% confidence bounds):
       p1 =     -0.8016  (-2.331, 0.7275)
       p2 =       14.93  (-15.22, 45.09)
       p3 =      -64.36  (-236.5, 107.8)
       p4 =       318.3  (49.04, 587.5)

Goodness of fit:
  SSE: 4.074e+004
  R-square: 0.3885
  Adjusted R-square: 0.1591
  RMSE: 71.36
```

2. Степеневі моделі (Power) (рис. 3.8).

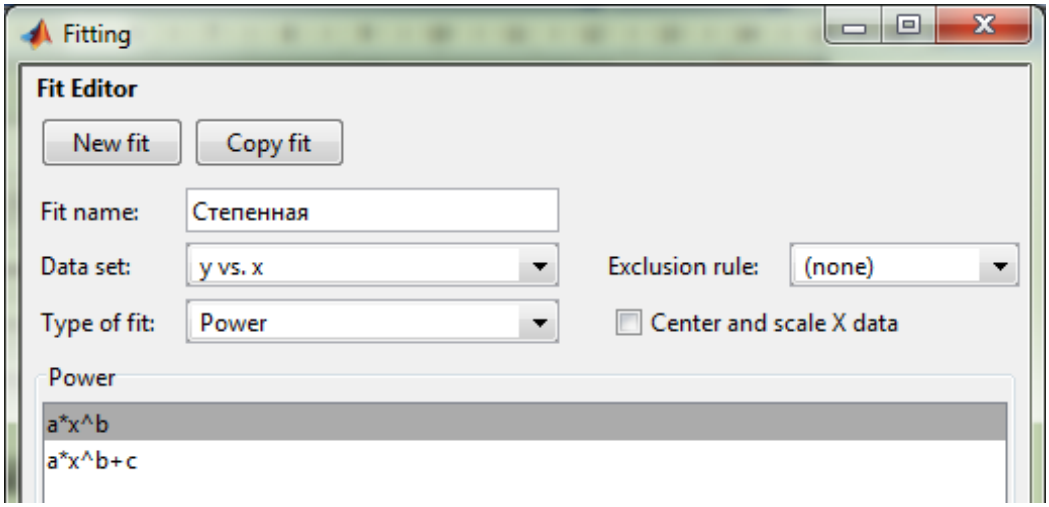


Рис. 3.8. **Вибір степеневої моделі діалогового вікна *Fitting***

```
General model Power1:
     f(x) = a*x^b
Coefficients (with 95% confidence bounds):
       a =       218.9  (118.9, 319)
       b =      0.1886  (-0.04665, 0.4239)

Goodness of fit:
  SSE: 4.895e+004
  R-square: 0.2653
  Adjusted R-square: 0.1918
  RMSE: 69.96
```

3. Експоненційні моделі (Exponential) (рис. 3.9).

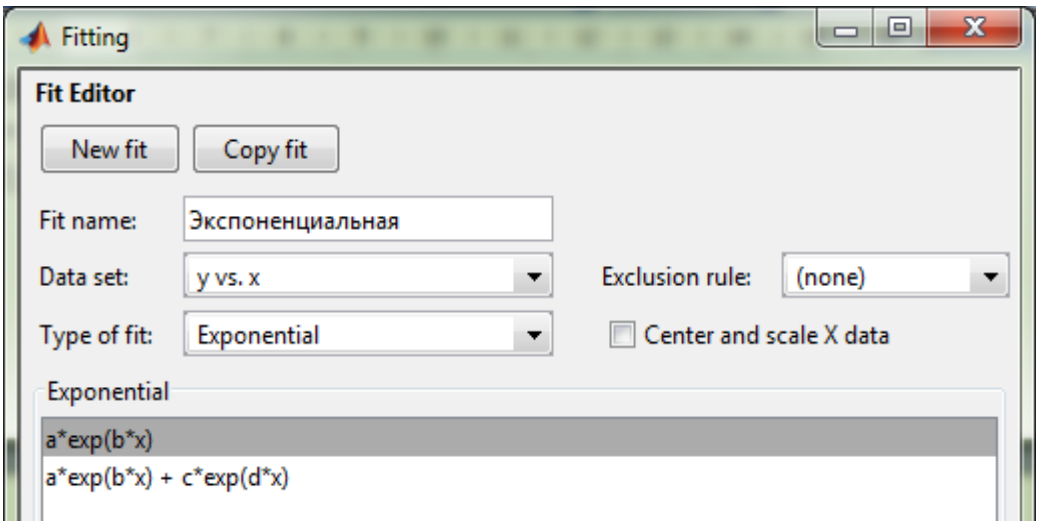


Рис. 3.9. **Вибір експоненційної моделі діалогового вікна *Fitting***

```
General model Exp1:
     f(x) = a*exp(b*x)

Coefficients (with 95% confidence bounds):
       a =       237.9  (153.8, 322)
       b =     0.03584  (-0.007869, 0.07956)

Goodness of fit:

  SSE: 4.931e+004
  R-square: 0.26
  Adjusted R-square: 0.1859
  RMSE: 70.22
```

4. Відрізки ряду Фур'є (*Fourier*) (рис. 3.10).

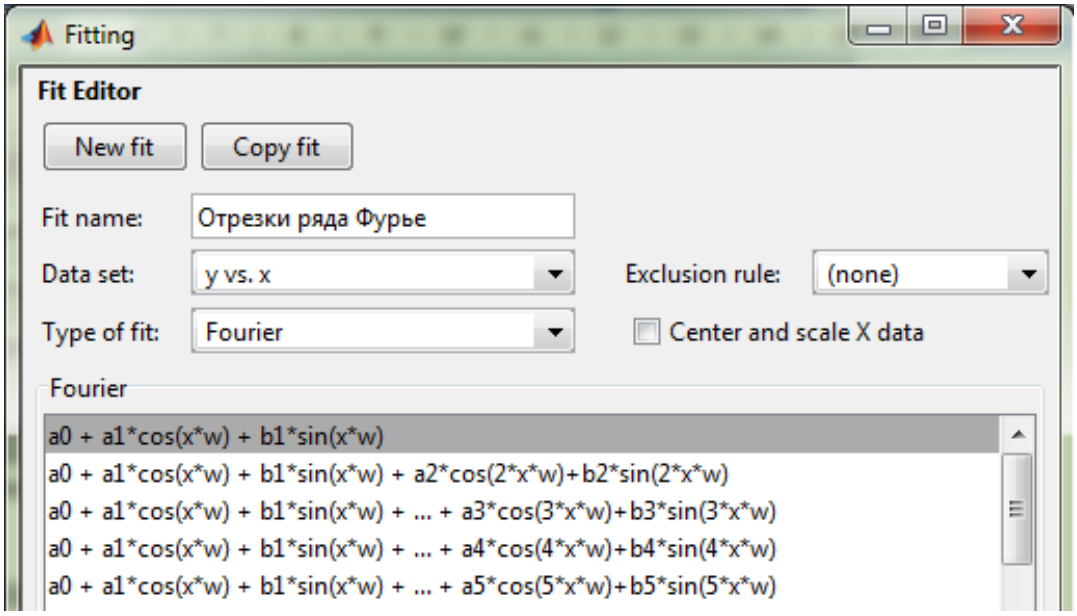


Рис. 3.10. **Вибір моделі "відрізки ряду Фур'є" діалогового вікна *Fitting***

```
General model Fourier1:
     f(x) =  a0 + a1*cos(x*w) + b1*sin(x*w)

Coefficients (with 95% confidence bounds):
       a0 =       302.2  (256.7, 347.7)
       a1 =       40.14  (-74.77, 155.1)
       b1 =      -60.98  (-141.5, 19.55)
       w =      0.5649  (0.3392, 0.7905)
```

```
Goodness of fit:

  SSE: 3.7e+004
  R-square: 0.4447
  Adjusted R-square: 0.2364
  RMSE: 68.01
```

5. Сума синусів (*Sum of Sin Functions*) (рис. 3.11).

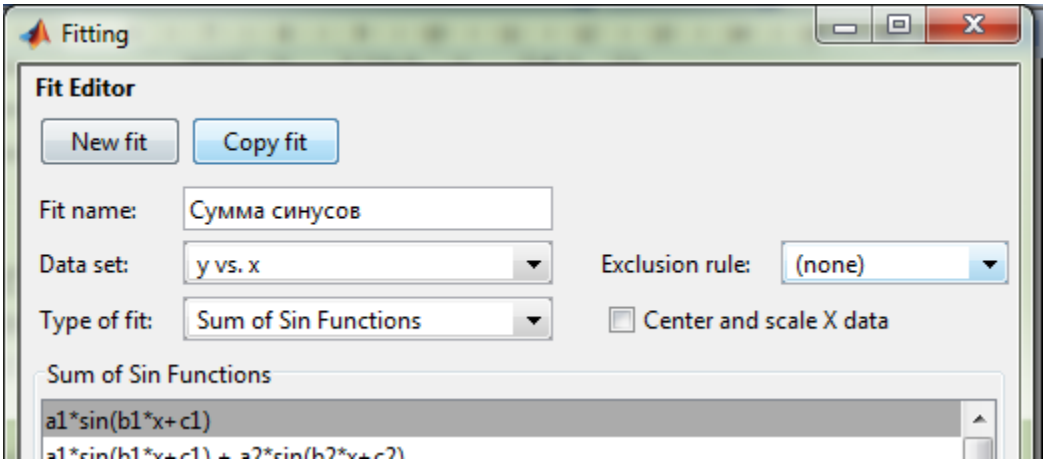


Рис. 3.11. **Вибір моделі "сума синусів" діалогового вікна *Fitting***

```
General model Sin1:
     f(x) =  a1*sin(b1*x+c1)

Coefficients (with 95% confidence bounds):

       a1 =        351.3  (158.1, 544.4)
       b1 =      0.07033  (-0.135, 0.2757)
       c1 =       0.6337  (0.1126, 1.155)


Goodness of fit:

  SSE: 4.804e+004
  R-square: 0.2789
  Adjusted R-square: 0.1187
  RMSE: 73.06
```

6. Гауссові моделі (*Gaussian*) (рис. 3.12).

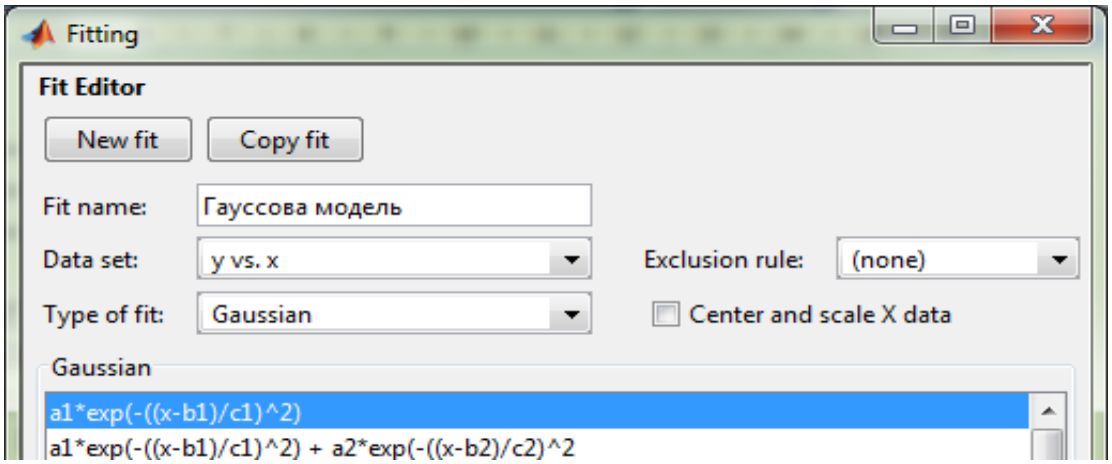


Рис. 3.12. **Вибір Гауссової моделі діалогового вікна *Fitting***

```
General model Gauss1:
     f(x) =  a1*exp(-((x-b1)/c1)^2)

Coefficients (with 95% confidence bounds):

       a1 =       347.1   (234.9, 459.4)
       b1 =       11.72   (-9.045, 32.48)
       c1 =       16.28   (-16.97, 49.53)

Goodness of fit:

  SSE: 4.758e+004
  R-square: 0.2858
  Adjusted R-square: 0.1271
  RMSE: 72.71
```

7. Модель Вейбулла (*Weibull*) (рис. 3.13).

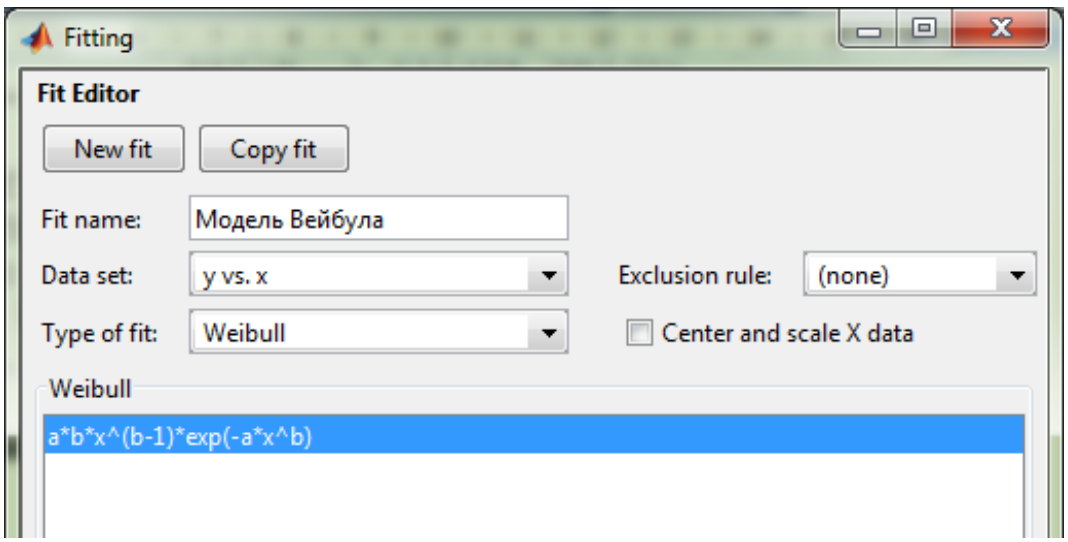


Рис. 3.13. **Вибір моделі Вейбулла діалогового вікна *Fitting***

```
General model Weibull:
     f(x) = a*b*x^(b-1)*exp(-a*x^b)

Coefficients (with 95% confidence bounds):

       a =     0.04694  (-249.3, 249.4)
       b =       3.059  (-4855, 4861)

Goodness of fit:

  SSE: 1.164e+006
  R-square: -16.47
  Adjusted R-square: -18.22
  RMSE: 341.2
```

8. Дробово-раціональні моделі (*Rational)* (рис. 3.14).

Ці моделі подають дробом, у чисельнику та знаменнику якого знаходяться поліноми до п'ятого степеня включно. Шуканими є коефіцієнти поліномів, що знаходяться у чисельнику та знаменнику дробу. Для вибору моделі цього типу в списку, що розкривається, ***Type of fit*** з'являються два списки для вибору ступеня чисельника та знаменника (рис. 3.14).

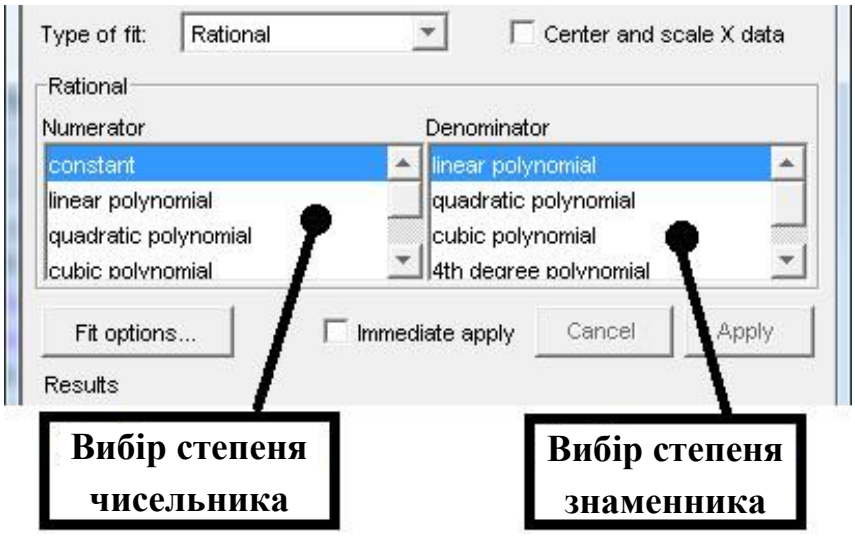


Рис. 3.14. **Діалогове вікно моделі *Rational***

На рис. 3.15 зображено підбір дробово-раціональної моделі, у чисельнику якої знаходиться константа, а в знаменнику – поліном першого степеня.

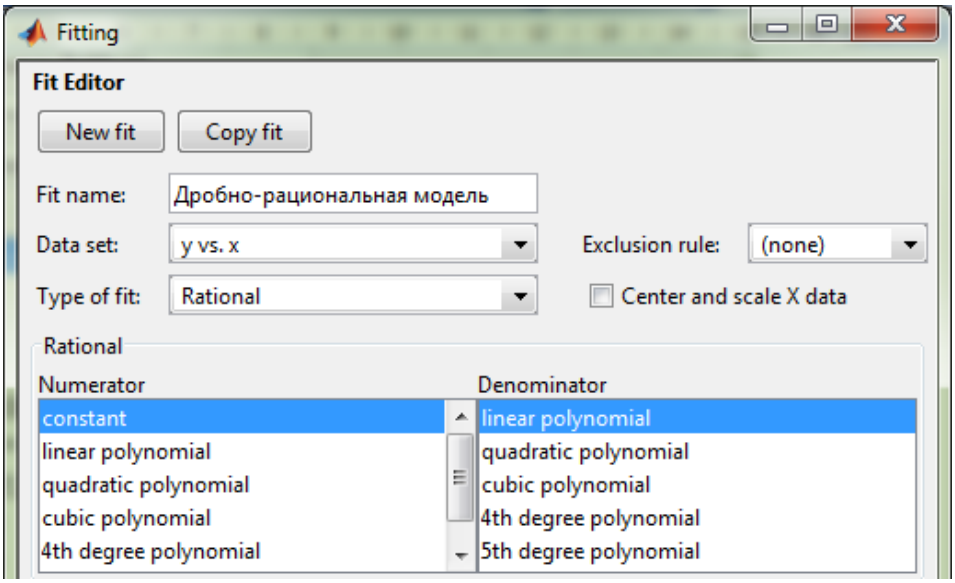


Рис. 3.15. **Вибір дробово-раціональної моделі діалогового вікна *Fitting***

```
General model Rat01:
     f(x) = (p1) / (x + q1)
Coefficients (with 95% confidence bounds):
       p1 =         95.1   (-254.9, 445.1)
       q1 =       -1.741   (-2.774, -0.7088)
Goodness of fit:
  SSE: 1.003e+006
  R-square: -14.06
  Adjusted R-square: -15.56
  RMSE: 316.7
```

Видно, що дана модель не придатна для побудови лінії регресії, однак вдалий підбір поліномів чисельника та знаменника дає гарну регресію:

```
General model Rat22:
   f(x) = (p1*x^2 + p2*x + p3) / (x^2 + q1*x + q2)

Coefficients (with 95% confidence bounds):

       p1 =        518.3   (-979.2, 2016)
       p2 =       -82.8   (-1.583e+004, 1.566e+004)
       p3 =       -6084   (-7.593e+004, 6.376e+004)
       q1 =        6.333   (-93.65, 106.3)
       q2 =      -33.02   (-361.9, 295.9)
```

```
Goodness of fit:

  SSE: 4.05e+004
  R-square: 0.3921
  Adjusted R-square: 0.0447
  RMSE: 76.07
```

Для проведення порівняльного аналізу різних параметричних моделей пакета *Curve Fitting Toolbox* треба вибрати найкращу модель, для чого об'єднати результати побудованих парних регресій в одній таблиці (табл. 3.3).

Таблиця 3.3

Порівняльний аналіз результатів роботи параметричних моделей пакету *Curve Fitting Toolbox*

Вид регресії	**Рівняння регресії**	**SSE (сума квадратів залишків)**	**R-square (коефіцієнт детермінації)**	**Adjusted R-square (корегований коефіцієнт детермінації)**	**RMSE (середнє квадратичне відхилення)**
Поліноміальна (кубічна) модель	$\tilde{y} = -0.8016x^3 + 14.93x^2 - 64.36x + 318.3$	$4.074 \cdot 10^4$	0.39	0.1591	71.36
Степенева модель	$\tilde{y} = 218.9 \cdot x^{0.1886}$	$4.895 \cdot 10^4$	0.27	0.1918	69.96
Експоненційна модель	$\tilde{y} = 237.9e^{0.03584x}$	$4.931 \cdot 10^4$	0.26	0.1859	70.22
Відрізки ряду Фур'є	$\tilde{y} = 302.2 + 40.14\cos 0.56x - 60.98\sin 0.56x$	$3.7 \cdot 10^4$	0.45	0.2364	68.01
Сума синусів	$\tilde{y} = 351.3\sin(0.07x + 0.63)$	$4.804 \cdot 10^4$	0.28	0.1187	73.06
Гауссова модель	$\tilde{y} = 347.1e^{-\left(\frac{x-11.72}{16.28}\right)^2}$	$4.758 \cdot 10^4$	0.29	0.1271	72.71
Модель Вейбулла	$\tilde{y} = 0.05 \cdot 3.06 \cdot x^{3.06-1}e^{-0.05x^{3.06}}$	$116.4 \cdot 10^4$	-16.47	-18.22	341,5
Дробово-раціональна модель	$\tilde{y} = \frac{518.3x^3 - 82.8x - 6084}{x^2 + 6.3x - 33.02}$	$4.05 \cdot 10^4$	0.39	0.0447	76.07

Усі рівняння регресії (крім моделі Вейбулла) досить добре описують вихідні дані. Деяку перевагу можна віддати відрізкам ряду Фур'є, кубічній та дробово-раціональній моделям, для яких значення коефіцієнта детермінації найбільше.

Усі моделі парної нелінійної регресії наведені на рис. 3.16.

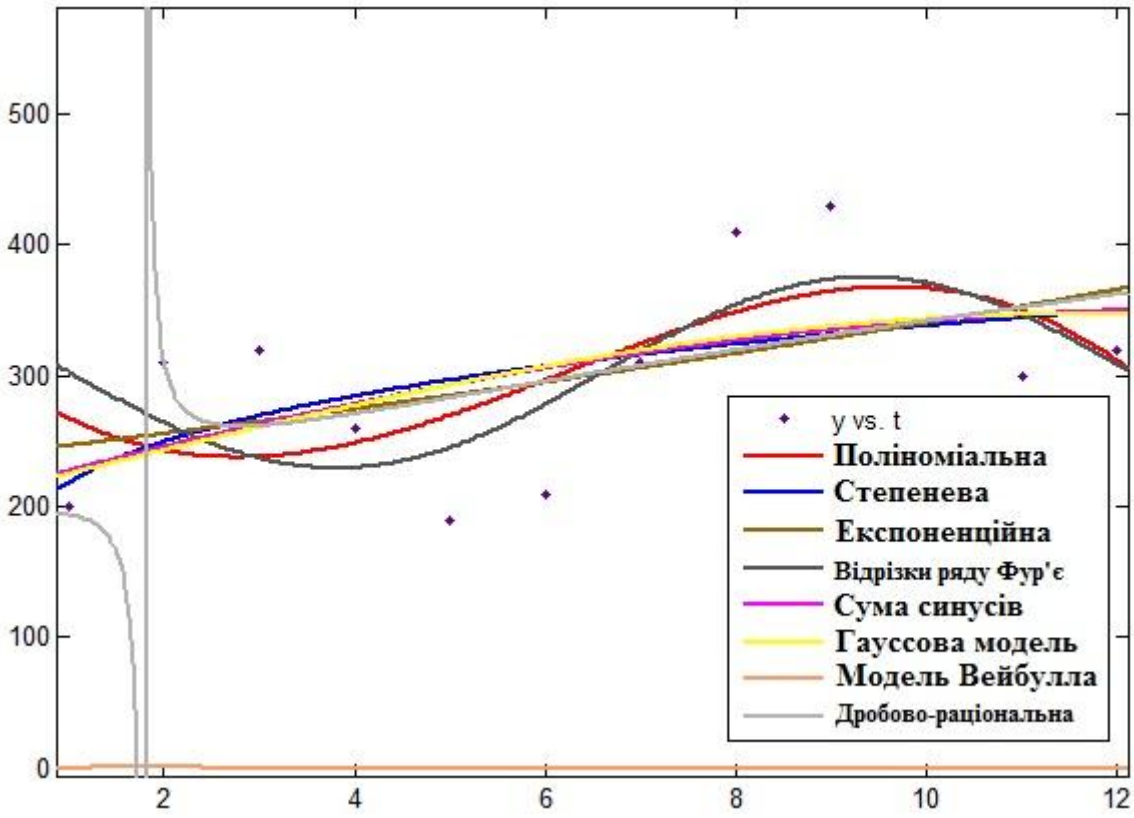


Рис. 3.16. **Нелінійні моделі парної регресії**

II. Непараметричні моделі

1. Інтерполяційні сплайни (*Interpolant*):

Linear – кусково-лінійне наближення (точки даних з'єднуються відрізками прямих).

nearest neibour – кусково-стала інтерполяція за найближчим сусідом.

cubic spline – інтерполяція даних кубічним сплайном. Створюється той самий сплайн, що будує функція *spline*, який входить у набір функцій MATLAB.

Shape-preserving – інтерполяція сплайном ерміта, тобто сплайном, що зберігає форму даних.

На рис. 3.17 зображено непараметричну модель "Інтерполяційний сплайн (*Interpolant*)".

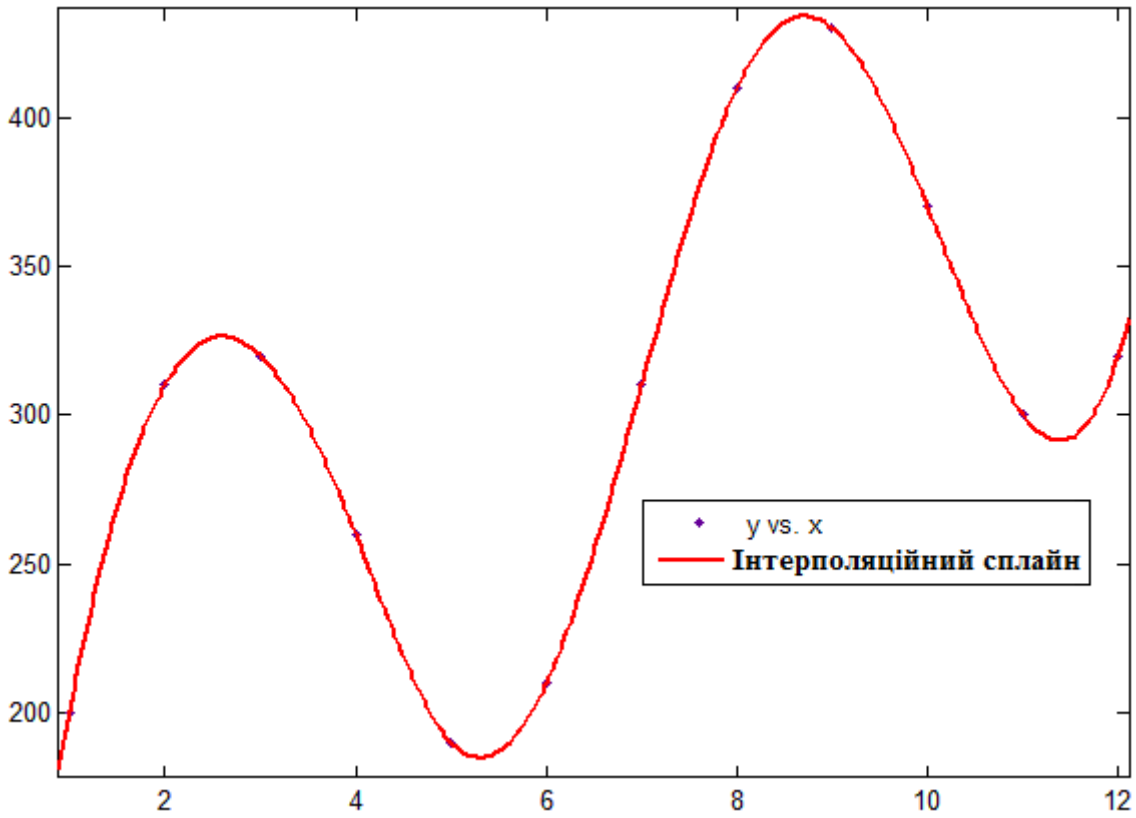


Рис. 3.17. **Графічне зображення моделі, отриманої за допомогою інтерполяційного сплайна**

2. Смузінг-сплайн (*Smoothing Spline*):

Значення згладжувального параметра задаються в діалоговому вікні *Fitting* (відповідні перемикачі, кнопки й область уведення з'являються після вибору ***Smoothing*** *Spline* у списку, що розкривається***, Type of*** *fit*), зображеного на рис. 3.18.

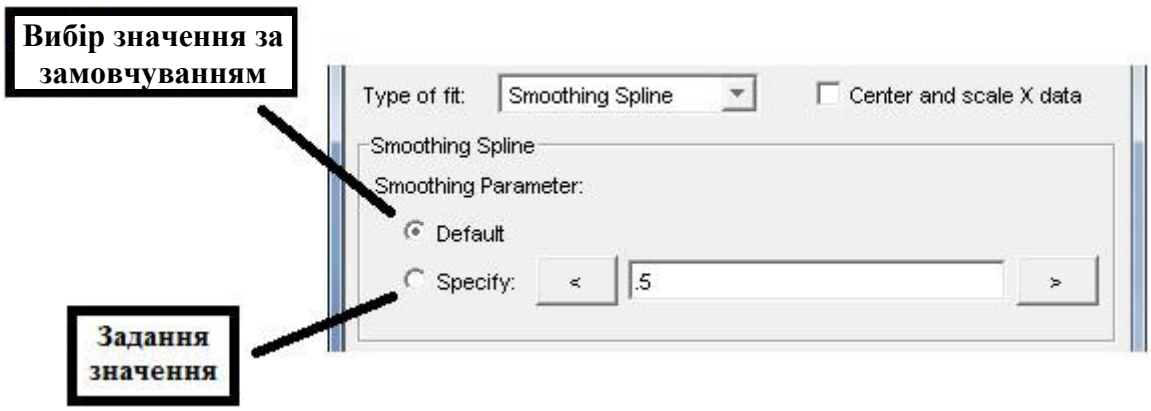


Рис. 3.18. **Діалогове вікно *Smoothing Spline***

На рис. 3.19 наведено побудовані моделі сплайнів для різних значень параметра *p*:

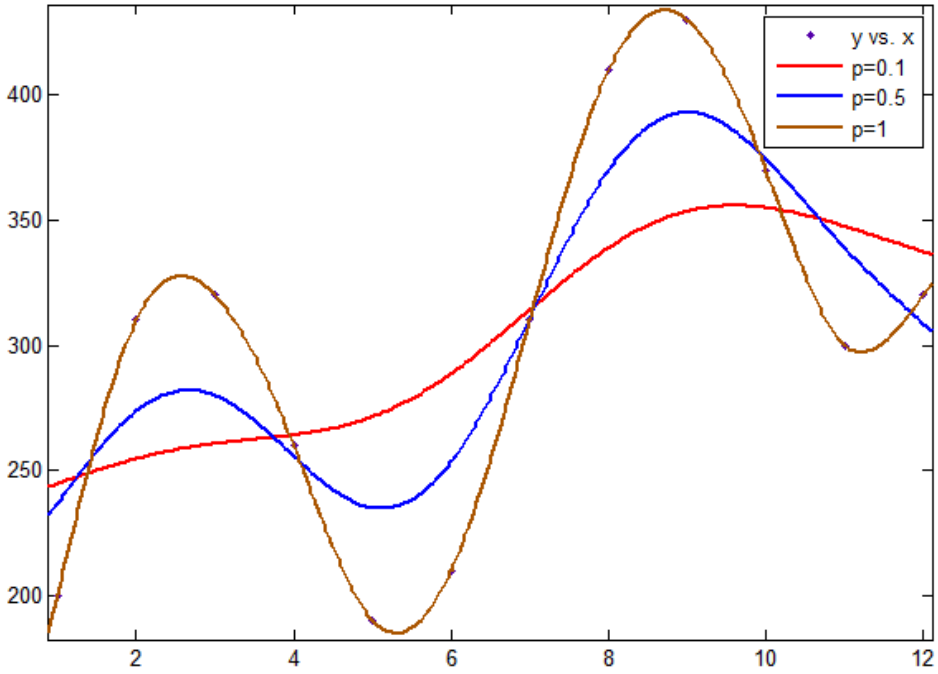


Рис. 3.19. **Наближення даних згладжувальними сплайнами**

Після того як результати були експортовані в робоче середовище (кнопка ***Save to workspace…***), з ними можна працювати. Це досягається за допомогою написання наступного програмного коду в М-файлі.

Фрагмент М-файла

```
%% Робота з експортованими даними:
p=coeffvalues(fittedmodel1)% коефіцієнти полінома
goodness1.sse % сума квадратів помилок
goodness1.rsquare % критерій R-квадрат
goodness1.dfe % число ступенів свобод
goodness1.adjrsquare % уточнений критерій R-квадрат
goodness1.rmse % корінь із середнього для квадрата помилки
```

Запитання для самоперевірки

1. Які існують види регресії?
2. Які існують види парної нелінійної регресії?
3. Які види залежностей можна звести функціональним перетворенням до лінійних?

4. Назвіть показники кореляції, що використовують при нелінійних співвідношеннях розглянутих ознак.

5. Який критерій використовують для оцінювання значущості рівнянь регресії?

Завдання для лабораторної роботи

У табл. 3.4 наведені дані (p_1 – кількість букв у повному імені, p_2 – кількість букв у прізвищі).

Таблиця 3.4

Вихідні дані задачі

Рік	**Фактичне кінцеве споживання домашніх господарств у. о., *y***	**Середньодушові грошові доходи населення (на місяць), у. о, *x***
1995	872 + p_1	515,9 - p_2
2000	3813 + p_1	2281,1 - p_2
2001	5014 - p_1	3062 + p_2
2002	6400 - p_1	3947,2 + p_2
2003	7708 + p_1	5170,4 - p_2
2004	9848 + p_1	6410,3 - p_2
2005	12455 - p_1	8111,9 + p_2
2006	15284 - p_1	10196 + p_2
2007	18928 + p_1	12602,7 - p_2
2008	23695 - p_1	14940,6 - p_2
2009	25151 + p_1	16856,9 + p_2

За даними з табл. 3.4 виконати наступні дії.

1. Знайти рівняння регресії, використовуючи вбудовані параметричні моделі: експоненційні, відрізки ряду Фур'є, сума синусів, Гауссові моделі, модель Вейбулла, статичні моделі, поліноміальні моделі, дробово-раціональні моделі.

2. Знайти рівняння регресії, використовуючи вбудовані непараметричні моделі: інтерполяційні та згладжувальні сплайни.

3. Провести дисперсійний аналіз.

4. Порівняти отримані моделі за допомогою коефіцієнта детермінації.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

Лабораторна робота 4
Багатофакторна лінійна регресійна модель

Мета роботи: набути компетентності з розроблення рівняння багатофакторної лінійної регресії за допомогою вбудованих функцій Matlab (*Statistic Toolbox*).

Основні задачі лабораторної роботи:

1. Знайти рівняння лінійної трьохфакторної регресії за допомогою стандартного МНК у матричній формі та з використанням функції `regress`. Порівняти отримані результати.

2. Перевірити значущість параметрів множинного рівняння регресії з використанням *t*-критерію Стьюдента.

3. Використовуючи функцію `corrcoef`, отримати матрицю парних коефіцієнтів кореляції, обчислити коефіцієнт множинної кореляції та детермінації.

4. Обчислити показники тісноти зв'язку факторів з результатом: частинні коефіцієнти еластичності, *β*-коефіцієнти, частинні коефіцієнти кореляції. Дати економічну інтерпретацію отриманих даних.

5. Побудувати модель у стандартизованих змінних.

6. Перевірити значущість отриманої моделі з використанням *F*-критерію Фішера.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

4.1. Основні поняття багатофакторної лінійної регресії

Основна мета множинної регресії – побудувати модель з великою кількістю факторів, визначивши вплив кожного з них окремо, а також сукупний їх вплив на результівний показник.

Виходячи з чіткої інтерпретації параметрів, найбільш широко використовуються в економіці лінійна та степенева залежності. Рівняння лінійної множинної регресії $y = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m + \varepsilon$ (рівняння "чистої" регресії). Коефіцієнти "чистої" регресії b_i характеризують середню зміну результату зі зміною відповідного фактора на одиницю за незмінним значенням інших факторів, закріплених на середньому рівні. У степеневій функції $\hat{y}_x = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot ... \cdot x_m^{b_m}$ коефіцієнти b_i (коефіцієнти еластичності) показують, на скільки відсотків змінюється в середньому результат із зміною відповідного

фактора на 1 % за незмінністю дії інших факторів. Цей вид рівняння регресії отримав найбільше поширення у виробничих функціях, у дослідженнях попиту та споживання.

Для моделювання залежностей в економіці використовуються й інші рівняння множинної регресії, які можна лінеаризувати:

$$\text{експоненційна} - \ y = e^{a+b_1x_1+b_2x_2+...+b_mx_m+\varepsilon}; \tag{4.1}$$

$$\text{гіперболічна} - \ y = \frac{1}{a + b_1x_1 + b_2x_2 + ... + b_mx_m}. \tag{4.2}$$

Для матричної форми рівняння регресії вектор коефіцієнтів регресії має вигляд:

$$B = \left(X^T X\right)^{-1} X^T Y, \tag{4.3}$$

де X^T – транспонована матриця X;

$C = \left(X^T X\right)^{-1}$ – обернена матриця.

Оцінювання ймовірності кожного з параметрів моделі здійснюється за допомогою t-критерію Стьюдента. Для кожного з параметрів моделі b_j значення t-критерію обчислюється за формулою:

$$t_{расч} = \frac{b_j}{S_\varepsilon \sqrt{c_{ij}}}, \tag{4.4}$$

де S_ε – стандартизоване (середнє квадратичне) відхилення рівняння регресії,

$S_\varepsilon = \sqrt{\dfrac{\sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right)^2}{n-m-1}}$;

c_{ij} – діагональні елементи матриці C.

Коефіцієнт регресії b_j вважається значущим, якщо обчислене значення t-критерію Стьюдента з (n – m – 1) ступенями свободи перевищує табличне, тобто $t_{расч} > t_{\alpha, n-m-1}$.

Для порівняльного оцінювання та відсіювання частини факторів знаходять матрицю парних коефіцієнтів кореляції (табл. 4.1), які вимірюють тісноту лінійного зв'язку кожного фактору з результівною ознакою та з кожним з інших факторів.

Таблиця 4.1

Матриця парних лінійних коефіцієнтів кореляції

	y	x_1	x_2	…	x_j	…	x_n
y	1	r_{yx_1}	r_{yx2}	…	r_{yx_j}	…	r_{yx_n}
x_1	r_{x_1y}	1	$r_{x_1x_2}$	…	$r_{x_1x_j}$	…	$r_{x_1x_n}$
x_2	r_{x_2y}	$r_{x_2x_1}$	1	…	$r_{x_2x_j}$	…	$r_{x_2x_n}$
…	…	…	…	…	…	…	…
x_i	r_{x_iy}	$r_{x_ix_1}$	$r_{x_ix_2}$	…	1	…	$r_{x_ix_n}$
…	…	…	…	…	…	…	…
x_n	r_{x_ny}	$r_{x_nx_1}$	$r_{x_nx_2}$	…	$r_{x_nx_j}$	…	1

де y – результівна ознака;

$x_1, x_2, \ldots, x_n$ – факторні ознаки;

r_{ij} – парний коефіцієнт кореляції між ознаками.

Показник множинної кореляції характеризує тісноту зв'язку розглянутого набору факторів з досліджуваною ознакою або оцінює тісноту спільного впливу факторів на результат.

Незалежно від форми зв'язку показник множинної кореляції може бути знайдений як індекс множинної кореляції:

$$R_{yx_1\ldots x_m} = \sqrt{1 - \frac{S_\varepsilon^2}{S_y^2}}, \qquad (4.5)$$

де S_y^2 – загальна дисперсія результівної ознаки;

S_ε^2 – залишкова дисперсія.

$0 \le R_{yx_1\ldots x_m} \le 1$, чим ближче його значення до 1, тим тісніший зв'язок результівної ознаки з усім набором досліджуваних факторів. З правильним вибором факторів у регресійний аналіз величина індексу множинної кореляції буде істотно відрізнятись від індексу кореляції парної залежності.

Однак класичний коефіцієнт множинної детермінації не завжди здатний визначити вплив на якість моделі регресії додаткової факторної змінної. Тому

поряд зі звичайним коефіцієнтом розраховують також і скоректований (*adjusted*) коефіцієнт множинної детермінації, у якому враховується кількість факторних змінних, включених у модель регресії:

$$R^2 = 1-(1-R^2)\frac{n-1}{n-m-1}, \qquad (4.6)$$

де n – кількість спостережень у вибірковій сукупності;

m – число параметрів, включених у модель регресії.

Для визначення рейтингу впливу факторів у моделі обчислюють регресійну модель в стандартизованих змінних. Усі формули регресійного аналізу в стандартизованих змінних набувають простішого вигляду. Якщо позначити $t_y = \frac{y-\bar{y}}{s_y}$, $t_{x_i} = \frac{x_i - \bar{x}_i}{s_{x_i}}$, тоді $t_y = \beta_1 t_1 + \beta_2 t_2 + ... + \beta_m t_m + \varepsilon$.

Коефіцієнти "чистої" регресії пов'язані зі стандартизованими коефіцієнтами регресії:

$$\beta_i = b_i \frac{s_{x_i}}{s_y},\ a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - ... - b_m\bar{x}_m. \qquad (4.7)$$

Стандартизовані коефіцієнти регресії показують, на скільки сигм (середньоквадратичних відхилень) зміниться в середньому результат, якщо відповідний фактор x_i зміниться на одну сигму з незмінним середнім рівнем інших факторів. Оскільки змінні центровані та нормовані, стандартизовані коефіцієнти регресії β_i порівнянні між собою і тому можна ранжувати фактори за силою їх впливу на результат.

У разі лінійної залежності ознак формула індексу кореляції може бути подана у вигляді:

$$R = \sqrt{\beta_{x_1} \cdot r_{yx_1} + \beta_{x_2} \cdot r_{yx_2} + ... + \beta_{x_m} \cdot r_{yx_m}} = \sqrt{\sum \beta_{x_i} \cdot r_{yx_i}}\ . \qquad (4.8)$$

Частинні коефіцієнти (індекси) кореляції характеризують тісноту зв'язку між результатом і відповідним фактором з усуненням впливу інших факторів, включених у рівняння регресії.

Порядок частинного коефіцієнта кореляції визначається кількістю чинників, вплив яких виключається. Наприклад, $r_{yx_1 \cdot x_2}$ – коефіцієнт частинної кореляції першого порядку. Коефіцієнти частинної кореляції більш високих

порядків можна визначити через коефіцієнти частинної кореляції більш низьких порядків за рекурентною формулою:

$$r_{yx_i \cdot x_1 x_2 \ldots x_m} = \frac{r_{yx_i \cdot x_1 x_2 \ldots x_{m-1}} - r_{yx_m \cdot x_1 x_2 \ldots x_{m-1}} \cdot r_{x_i x_m \cdot x_1 x_2 \ldots x_{m-1}}}{\sqrt{\left(1 - r^2_{yx_m \cdot x_1 x_2 \ldots x_{m-1}}\right)\left(1 - r^2_{x_i x_m \cdot x_1 x_2 \ldots x_{m-1}}\right)}}. \quad (4.9)$$

З метою розширення можливостей змістовного аналізу моделі регресії використовуються частинні коефіцієнти еластичності, які визначаються за формулою:

$$Э_i = b_i \frac{\bar{x}_i}{\bar{y}}. \quad (4.10)$$

Частинний коефіцієнт еластичності показує, на скільки відсотків у середньому змінюється ознака-результат y зі збільшенням ознаки-фактору x_i на 1 % від свого середнього рівня з фіксованим положенням інших факторів моделі.

Значущість рівняння множинної регресії в цілому оцінюється за допомогою F-критерію Фішера. Перевіряється гіпотеза H_0 про статистичну значущість коефіцієнта детермінації ($H_0 : R^2 = 0$). Для перевірки даної гіпотези використовується F-статистика:

$$F_p = \frac{S^2_{y_x}}{S^2_\varepsilon} = \frac{R^2}{1 - R^2} \cdot \frac{n - 1 - m}{m}, \quad (4.11)$$

де $S^2_{y_x}$ – факторна сума квадратів на одну ступінь свободи;

S^2_ε – залишкова сума квадратів на одну ступінь свободи;

R^2 – коефіцієнт детермінації;

m – число параметрів у змінних (у лінійній регресії збігається з числом включених у модель факторів);

n – кількість спостережень.

Показники F і R^2 водночас дорівнюють або ні нулю. Якщо F = 0, то R^2 = 0 лінія регресії $y = \bar{y}$ є найкращою за МНК. Отже, змінна y лінійно не залежить від $x_1, x_2, ..., x_m$. Для перевірки нульової гіпотези за заданим рівнем значущості α за таблицями критичних точок розподілу Фішера знаходять

критичне значення $F_{кр} = F_{\alpha}(m, n-m-1)$. Нульова гіпотеза відхиляється, якщо $F > F_{кр}$. Це рівнозначно тому, що $R^2 > 0$, тобто R^2 статистично значущий.

4.2. Теоретичні відомості про функції *Matlab*, які використовуються в даній лабораторній роботі

REGRESS **Множинна лінійна регресія**

Синтаксис

```
b = regress(y,X)
[b,bint,r,rint,stats] = regress(y,X)
[b,bint,r,rint,stats] = regress(y,X,alpha)
```

Опис

`b = regress(y,X)` функція призначена для обчислення точкових оцінок коефіцієнтів лінійного рівняння регресії *b*. Обчислення точкових оцінок коефіцієнтів виконується методом найменших квадратів з рівняння лінійної моделі $y = \beta X + \varepsilon$, де *y* – вектор значень залежної змінної; β – вектор коефіцієнтів лінійної моделі; *X* – матриця значень незалежних змінних; ε – вектор випадкових факторів, розподілених за нормальним законом з нульовим математичним очікуванням і дисперсією *S*.

Розмірності векторів значень залежної змінної *y* і випадкових факторів ε – $n \times 1$, де *n* – кількість спостережень. Розмірність матриці *X* дорівнює $n \times p$, де *p* – кількість незалежних змінних. Стовпці матриці *X* відповідають незалежним змінним, рядка – спостереженням. Розмірність вектора коефіцієнтів лінійної регресійної моделі дорівнює $p \times 1$. Коефіцієнти множинної лінійної регресійної моделі у векторі *b* розташовуються за зростанням ступеня незалежних змінних.

`[b,bint,r,rint,stats] = regress(y,X)` функція повертає: `b` – вектор точкових оцінок коефіцієнтів лінійного рівняння регресії, `bint` – матрицю інтервальних оцінок параметрів лінійної регресії, `r` – вектор залишків, `rint` – матрицю 95 % довірчих інтервалів залишків, `stats` – структуру, що містить значення статистики R^2 з відповідними їй *F*-статистикою та рівнем значущості *p* для регресійної моделі.

Розмірність матриці *bint* становить $p \times 2$, де перший стовпець матриці задає нижню межі 95 % довірчого інтервалу, другий – верхню межу 95 % довірчого інтервалу. Кількість елементів вектора *r* дорівнює *n*. Розмірність матриці *rint* дорівнює $n \times 2$, де перший та другий стовпці використовуються для задання нижньої та верхньої меж 95 % довірчого інтервалу за кожним з *n* спостережень.

[b,bint,r,rint,stats] = regress(y,X,alpha) – вхідний параметр *alpha* дозволяє задати величину рівня значущості. Рівень значущості використовується для обчислення меж довірчих інтервалів *bint* і *rint* з довірчою ймовірністю *100·(1-alpha)%*. Значення *alpha=0.2* буде відповідати 80 % меж довірчих інтервалів *bint* і *rint*.

MEAN **Визначення середніх значень елементів масиву**

Синтаксис:

mx = mean(X)

Опис:

Функція *mx = mean(X)* у випадку одномірного масиву повертає арифметичне середнє елементів масиву; у випадку двовимірного масиву – це рядок – вектор – рядок, що містить арифметичне середнє елементів кожного стовпця. Таким чином, *mean*(*mean*(*X*)) – це арифметичне середнє (математичне сподівання) елементів масиву, що збігається зі значенням *mean*(*X*(:)).

4.3. Розв'язування типової задачі в середовищі *Matlab*

Задача. Потрібно побудувати статистичну залежність вартості квартири від трьох факторів, наведених у табл. 4.2 у вигляді множинної регресії.

Таблиця 4.2

Вихідні відомості задачі

Вартість квартири, тис. дол., *Y*	**Загальна площа, м^2, *X*1**	**Житлова площа квартири, м^2, *X*2**	**Відстань до метро, хвилин пішки, *X*3**
1	2	3	4
16	80	84	3
22	62	37	8
23	69.7	42	18

Закінчення табл. 4.2

1	2	3	4
19.8	79	80.3	28
34	96.4	88	8
24.8	90	64	8
27.3	102	66	7
41	87	86.8	10
31	114.8	74	10
38.6	114.3	74.7	8
46	90	62	8
38	116	81	10
42.7	107	78.8	10
27	93	66	18
78	176	129	10
38	96	69.4	8
23.8	92	72.8	10
68	176	110	20
23	74	49	18
48.8	106	73.7	10
34	88	61.7	3
23	74	48.8	10
26.8	74.7	80.8	10
37	118	76	8
30	92	62	18
43	110	79.8	8

Для початку роботи необхідно створити новий файл. Для цього з меню *File* вибрати опцію *New*, а потім M-File. У вікні редактора М-файлів з використанням функції `dlmread`, отримано масив з файла даних. Також необхідно задати кількість спостережень (змінна *n*) і кількість факторів (змінна *m*). Необхідно звернути увагу, що в матрицю *D* додатково введений стовпець одиниць.

Фрагмент М-файла

```
%% Лінійна множинна регресія
clc % Очистити командне вікно
clear all
%% Отримати дані з файла
X = dlmread('lab4.txt');
```

```
x1 = X(1,:)';% значення x1
x2 = X(2,:)';% значення x2
x3 = X(3,:)';% значення x3
y = X(4,:)';% значення y
n = 26; % кількість спостережень
m = 3; % кількість факторів
D = [ones(n,1) x1 x2 x3];% формування вхідного масиву даних
```

Використовуючи навички й уміння роботи з масивами в середовищі Matlab, легко знайти рівняння лінійної множинної регресії за допомогою стандартного МНК й у матричній формі:

Фрагмент М-файла

```
%% Множинна лінійна регресія
% 1. Метод найменших квадратів:
A = ones(4);
A(1,1) = n;
A(1,2) = sum(x1); A(2,1) = A(1,2);
A(1,3) = sum(x2); A(3,1) = A(1,3);
A(1,4) = sum(x3); A(4,1) = A(1,4);
A(2,3) = sum(x1.*x2); A(3,2) = A(2,3);
A(2,4) = sum(x1.*x3); A(4,2) = A(2,4);
A(3,4) = sum(x2.*x3); A(4,3) = A(3,4);
A(2,2) = sum(x1.^2);
A(3,3) = sum(x2.^2);
A(4,4) = sum(x3.^2);
B = [sum(y); sum(x1.*y); sum(x2.*y); sum(x3.*y)];
b = inv(A)*B
% 2. Матрична форма:
b = inv(D'*D)*D'*y
```

Використовуючи вбудовану функцію *regress*, знайти рівняння множинної лінійної регресії.

Фрагмент М-файла

```
% 3. Використовуючи вбудовану функцію regress:
[b,bint,r,rint,stats] = regress(y,D,0.05);
fprintf('Уравнение линейной парной регрессии:')
y_p = subs(sym('a + b1*x1 + b2*x2 + b3*x3'), {'a', 'b1', 'b2',
'b3'}, [b(1) b(2) b(3) b(4)])
```

Для даної моделі всі три способи дають однакові значення параметрів моделі:

```
>>b =

   -9.6937
    0.4358
    0.0407
   -0.1532

>>b =

   -9.6937
    0.4358
    0.0407
   -0.1532

>>b =

   -9.6937
    0.4358
    0.0407
   -0.1532


>> Рівняння лінійної парної регресії
y_p = 0.4358*x1 + 0.0407*x2 - 0.1532*x3 - 9.6937
```

Перевірка значущості отриманих коефіцієнтів рівняння регресії здійснюється за допомогою *t*-розподілу Стьюдента.

Фрагмент М-файла

```
%% Перевірка значущості параметрів множинного рівняння регресії
t_tabl = 2.0739;% Ттабл (n-m-1;alfa) = (22;0.05) = 2.0739
C = inv(D'*D);
tb1 = b(2)/(sqrt(sum(r.^2) / 22)*sqrt(C(2,2)))
if abs(tb1) > t_tabl
    'Коефіцієнт регресії b1 можна вважати надійним'
else
    'Коефіцієнт регресії b1 не є надійним'
end
tb2 = b(3)/(sqrt(sum(r.^2) / 22)*sqrt(C(3,3)))
```

```
if abs(tb2) > t_tabl
    'Коефіцієнт регресії b2 можна вважати надійним'
else
    'Коефіцієнт регресії b2 не є надійним'
end
tb3 = b(4)/(sqrt(sum(r.^2) / 22)*sqrt(C(4,4)))
if abs(tb3) > t_tabl
    'Коефіцієнт регресії b3 можна вважати надійним'
else
    'Коефіцієнт регресії b3 не є надійним'
end
```

```
>>tb1 =

    4.8233

    Коефіцієнт регресії b1 можна вважати надійним

>>tb2 =

    0.3178

    Коефіцієнт регресії b2 не є надійним

>>tb3 =

   -0.5927

    Коефіцієнт регресії b3 не є надійним
```

Як видно з роботи програми, статистична значущість коефіцієнтів регресії b_2 і b_3 не підтверджується.

Використовуючи функцію *corrcoef*, отримати матрицю парних коефіцієнтів кореляції *R*. Використовуючи структуру *stats*, визначити значення коефіцієнта детермінації, а коефіцієнт множинної кореляції розрахувати, використовуючи формулу (4.5).

Фрагмент М-файла

```
fprintf('Матриця парних коефіцієнтів кореляції:')
R = corrcoef([y x1 x2 x3])
% коефіцієнт множинної кореляції
fprintf('коефіцієнт множинної кореляції:')
kor = sqrt(stats(1))
```

```
% коефіцієнт множинної детермінації
fprintf('коефіцієнт множинної детермінації:')
R2 = stats(1)
```

```
>> Матриця парних коефіцієнтів кореляції:

R =

    1.0000    0.8761    0.7279   -0.0387
    0.8761    1.0000    0.8079    0.0258
    0.7279    0.8079    1.0000    0.0026
   -0.0387    0.0258    0.0026    1.0000

>> Коефіцієнт множинної кореляції:

kor =

    0.8788

>> Коефіцієнт множинної детермінації:

R2 =

    0.7723
```

Використовуючи формули (4.8) – (4.10) обчислимо показники тісноти зв'язку факторів з результатом: частинні коефіцієнти еластичності, β-коефіцієнти, частинні коефіцієнти кореляції.

Фрагмент М-файла

```
%% Частинні коефіцієнти еластичності:
E1 = b(2)*mean(x1)/mean(y)
E2 = b(3)*mean(x2)/mean(y)
E3 = b(4)*mean(x3)/mean(y)
% Beta- коефіцієнти:
B = R(2:end, 1);% виріжемо з матриці R перший стовпець
A = R(2:end,2:end);% виріжемо з матриці R елементи
bet = inv(A)*B
% частинні коефіцієнти кореляції:
t_tabl = 2.069; % Ттабл (n-m-1;alfa/2) = (22;0.025) = 2.069
k = 1 % кількість фіксованих факторів
ryx1x2 = (R(1,2)-R(1,3)*R(2,3))/(sqrt(1-R(1,3)^2)*sqrt(1-R(2,3)^2))
ryx1x3 = (R(1,2)-R(1,4)*R(2,4))/(sqrt(1-R(1,4)^2)*sqrt(1-R(2,4)^2))
```

```
ryx2x1 = (R(1,3)-R(1,2)*R(3,2))/(sqrt(1-R(1,2)^2)*sqrt(1-R(3,2)^2))
ryx2x3 = (R(1,3)-R(1,4)*R(3,4))/(sqrt(1-R(1,4)^2)*sqrt(1-R(3,4)^2))
ryx3x1 = (R(1,4)-R(1,3)*R(4,2))/(sqrt(1-R(1,3)^2)*sqrt(1-R(4,2)^2))
ryx3x2 = (R(1,4)-R(1,3)*R(4,3))/(sqrt(1-R(1,3)^2)*sqrt(1-R(4,3)^2))
```

```
>>E1 =
    1.2420

>>E2 =
    0.0853

>>E3 =
   -0.0486

>>bet =
    0.8333
    0.0549
   -0.0603

>>ryx1x2 =
    0.7127
```

Частинний коефіцієнт кореляції ryx1x2 статистично значущий

```
>>ryx1x3 =
    0.8780
```

Частинний коефіцієнт кореляції ryx1x3 статистично значущий

```
>>ryx2x1 =
    0.0710
```

Частинний коефіцієнт кореляції ryx2x1 статистично не значущий

```
>>ryx2x3 =
    0.7285
```

```
Частинний коефіцієнт кореляції ryx2x3 статистично
значущий

>>ryx3x1 =
   -0.0839

Частинний коефіцієнт кореляції ryx3x1 статистично не
значущий

>>ryx3x2 =
   -0.0592

Частинний коефіцієнт кореляції ryx3x2 статистично не
значущий
```

Частинний коефіцієнт еластичності $|E1| > 1$. Отже, фактор істотно впливає на результівну ознаку y.

Частинний коефіцієнт еластичності $|E2| < 1$. Отже, вплив фактора на результівну ознаку y незначний.

Частинний коефіцієнт еластичності $|E3| < 1$. Отже, вплив фактора на результівну ознаку y незначний.

Використовуючи попередні значення β-коефіцієнтів, побудувати модель у стандартизованих змінних.

Фрагмент М-файла

```
%% Стандартизована форма рівняння
fprintf('Стандартизована форма рівняння регресії має вигляд:')
t_y = subs(sym('b1*t1 + b2*t2 + b3*t3'),{'b1','b2','b3'},[bet(1)
bet(2) bet(3)])
```

```
>> Стандартизована форма рівняння регресії має вигляд:

    t_y = 0.8333*t1 + 0.0549*t2 - 0.0603*t3
```

Оцінювання значущості рівняння множинної регресії здійснюють шляхом перевірки гіпотези про дорівненість нуля коефіцієнта детермінації, розрахованого за даними генеральної сукупності. Для її перевірки використовують F-критерій Фішера (його значення так само містить структура `stats`).

Фрагмент М-файла

```
%% Значущість моделі (F-критерій Фішера)
F_tabl = 3.05;
F_mod(1) = stats(2)
if abs(F_mod) > F_tabl
    fprintf('Коефіцієнт детермінації статистично значущий; рівняння регресії статистично надійне')
else
    fprintf('Коефіцієнт детермінації не є статистично значущим; рівняння регресії не є статистично надійним')
end
```

```
>>F_mod =
   24.8706
```

Коефіцієнт детермінації статистично значущий; рівняння регресії статистично надійне.

Запитання для самоперевірки

1. У чому суть специфікації множинної регресійної моделі?
2. Як оцінюють параметри множинної лінійної регресійної моделі?
3. Яка мета побудови регресійної моделі в стандартизованих змінних?
4. Який зв'язок між коефіцієнтами множинної лінійної "чистої" регресії і в стандартизованих змінних?
5. Як інтерпретуються β-коефіцієнти?
6. Як обчислюють індекс множинної лінійної кореляції?
7. Яке основне призначення коефіцієнтів приватної кореляції?
8. Як використовувати критерій Фішера для перевірки загальної якості рівняння регресії?
9. Які існують методи перевірки статистичної значущості коефіцієнтів рівняння регресії?
10. Чи існує зв'язок між методами перевірки статистичної значущості коефіцієнтів рівняння регресії?

Завдання до лабораторної роботи

Задача. У табл. 4.3 містяться дані про нерухомість в місті.

Таблиця 4.3

Вихідні дані задачі про нерухомість

Вартість квартири, тис. дол., *Y*	**Загальна площа, м^2, *X*1**	**Житлова площа квартири, м^2, *X*2**	**Відстань до метро, хв. пішки, *X*3**
16	80 + p_1	84 – p_2	3
22	62 – p_1	37 – p_2	8
23	69.7 + p_1	42 + p_2	18
19.8	79 + p_1	80.3 – p_2	28
34	96.4 – p_1	88 – p_2	8
24.8	90 – p_1	64 + p_2	8
27.3	102 + p_1	66 + p_2	7
41	87 – p_1	86.8 – p_2	10
31	114.8 + p_1	74 – p_2	10
38.6	114.3 + p_1	74.7 – p_2	8
46	90 + p_1	62 + p_2	8
38	116 – p_1	81 – p_2	10
42.7	107 + p_1	78.8 + p_2	10
27	93 + p_1	66 + p_2	18
78	176 – p_1	129 – p_2	10
38	96 – p_1	69.4 + p_2	8
23.8	92 + p_1	72.8 + p_2	10
68	176 + p_1	110 – p_2	20
23	74 + p_1	49 + p_2	18
48.8	106 – p_1	73.7 + p_2	10
34	88 + p_1	61.7 + p_2	3
23	74 + p_1	48.8 + p_2	10
26.8	74.7 + p_1	80.8 – p_2	10
37	118 – p_1	76 + p_2	8
30	92 – p_1	62 + p_2	18
43	110 – p_1	79.8 – p_2	8

Знайти рівняння лінійної трьохфакторної регресії за допомогою стандартного МНК у матричній формі й з використанням функції `regress`. Порівняти отримані результати.

1. Перевірити значущість параметрів множинного рівняння регресії з використанням *t*-критерію Стьюдента.

2. Використовуючи функцію `corrcoef`, отримати матрицю парних коефіцієнтів кореляції, обчислити коефіцієнт множинної кореляції та детермінації.

3. Обчислити показники тісноти зв'язку факторів з результатом: частинні коефіцієнти еластичності, β-коефіцієнти, частинні коефіцієнти кореляції. Дати економічну інтерпретацію отриманих даних.

4. Побудувати модель у стандартизованих змінних.

5. Перевірити значущість отриманої моделі з використанням F-критерію Фішера.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

Лабораторна робота 5
Мультиколінеарність, її наслідки та методи усунення

Мета роботи: мати уявлення щодо мультиколінеарності і її негативного впливу на статистичну якість моделі множинної регресії, освоїти способи визначення мультіколінеарності факторів моделі, навчитися усувати мультіколінеарність різними способами.

Основні задачі лабораторної роботи:

1. Побудувати модель множинної регресії, здійснити кореляційно-регресійний аналіз моделі засобами *StatisticToolbox*.

2. Визначити наявність або відсутність мультиколінеарності факторів.

3. За наявності мультіколінеарності усунути її за допомогою вбудованих функцій *MATLAB*.

4. Побудувати модель множинної регресії, використовуючи тільки значущі фактори.

5. Провести порівняльний аналіз моделей та дати економічну інтерпретацію отриманих результатів.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

5.1. Мультиколінеарність. Виявлення. Методи усунення

Серйозною проблемою у процесі побудови моделей множинної лінійної регресії за МНК є мультиколінеарність – лінійний взаємозв'язок двох або декількох пояснювальних змінних. Мультиколінеарність може бути проблемою лише у разі множинної регресії.

Причиною виникнення мультиколінеарності в економічних дослідженнях є існування співвідношень між пояснювальними змінними. Це стосується регресії, побудованої як на результатах одночасних обстежень, так і за даними, отриманими з часових рядів.

Зазвичай виділяються такі наслідки мультиколінеарності:

1) значні дисперсії (стандартні помилки) оцінок. Це ускладнює знаходження істинних значень визначених величин і розширює інтервальні оцінки, погіршуючи їх точність;

2) зменшуються *t*-статистики коефіцієнтів, що може призвести до невиправданого висновку про суттєвість впливу відповідної пояснювальної змінної на залежну змінну;

3) оцінки коефіцієнтів з МНК і їх стандартні помилки стають дуже чутливими до найменших змін даних, тобто вони стають нестійкими;

4) ускладнюється визначення внеску кожної зі змінних, які пояснюють дисперсію, що пояснюється рівнянням регресії залежної змінної;

5) можливе отримання неправильного знака у коефіцієнта регресії.

Існує кілька ознак, за якими може бути встановлена наявність мультиколінеарності:

1) коефіцієнт детермінації R^2 досить високий, але деякі з коефіцієнтів регресії статистично незначущі, тобто вони мають низькі *t*-статистики;

2) парна кореляція між малозначущими пояснювальними змінними досить висока. За великої їх кількості доцільнішим є використання частинних коефіцієнтів кореляції;

3) високі частинні коефіцієнти кореляції;

4) сильна допоміжна (додаткова регресія), тобто будь-яка з пояснювальних змінних є лінійною комбінацією інших пояснювальних змінних.

Єдиного методу усунення мультиколінеарності не існує, але є рекомендації, які часто використовуються.

1. *Вилучення змінної з моделі.* Цей метод полягає в тому, що пояснювальні змінні, які високо корелюють, виключаються з регресії, і вона заново оцінюється. Практикою доведено, що якщо $|r_{ij}| > 0{,}8$, то одну зі змінних слід виключити. Яку саме змінну необхідно виключити, визначають на підставі економічного аналізу залежної змінної, оскільки можна допустити помилку специфікації. Наприклад, у ході дослідження попиту на деякий товар у якості пояснювальної змінної можна взяти ціну даного товару та ціну його замінника, які часто

корелюють один з одним. Виключення з моделі ціни замінника може викликати помилку специфікації. Внаслідок цього можна отримати зміщені оцінки, і висновки будуть необґрунтованими.

2. *Отримання додаткових даних або нової вибірки.* Мульти-колінеарність прямо залежить від вибірки, тому у процесі переходу до іншої вибірки мультиколінеарность буде відсутня або не буде значною. У такому випадку для зменшення мультиколінеарності достатньо збільшити обсяг вибірки. Але й тут можуть виникнути проблеми, пов'язані з отриманням нової вибірки або розширенням старої.

3. *Зміна специфікації моделі*, тобто змінюється форма моделі або додаються пояснювальні змінні, не враховані в первісній моделі, але такі, що істотно впливають на залежну змінну.

4. *Використання попередньої інформації про деякі параметри*. Іноді на основі раніше побудованих регресійних рівнянь або проведених економічних досліджень формується уявлення про величину або співвідношення двох або декількох коефіцієнтів регресії. Тут можуть виникнути труднощі, обумовлені способом отримання попередньої інформації та малою ймовірністю того, що виділений коефіцієнт регресії буде однаковим у різних моделей.

5. *Перетворення змінних.* Часто у ході розв'язування економічних задач для усунення проблеми мультиколінеарності використовують метод перетворення змінних. Наприклад, емпіричне рівняння регресії має вигляд: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$, причому змінні x_1, x_2 корелюють між собою. Тоді можна визначити регресійні залежності відносних величин:

$$\frac{\hat{y}}{x_1} = b_0 + b_1 \frac{x_2}{x_1} \text{ або } \frac{\hat{y}}{x_2} = b_0 + b_1 \frac{x_1}{x_2}. \qquad (5.1)$$

Цілком імовірно, що в цих моделях мультиколінеарність буде відсутня. Можливі й інші перетворення, що схожі до наведених.

6. Вирішенням проблеми мультиколінеарності може стати *коригування самого математичного методу оцінювання параметрів регресійної моделі.*

7. *Метод головних компонент.* Результатом застосування методу головних компонент на вихідній системі пояснювальних змінних є нові змінні, які є лінійною комбінацією вихідних.

5.2. Теоретичні відомості про функції *MATLAB*, які використовуються в даній лабораторній роботі

REGSTATS

Обчислення параметрів і діагностика множинної регресійної моделі

Синтаксис

```
regstats(responses,DATA,'model')
stats = regstats(responses,DATA,'model','whichstats')
```

Опис

regstats(*responses*, *DATA*,'*model*') – функція призначена для обчислення параметрів множинної регресійної моделі для вектора значень залежної змінної *responses*, матриці незалежних змінних *DATA*, регресійної моделі '*model*'. Функція відображує графічне вікно з набором статистик, які служать для оцінювання якості множинної регресійної моделі (рис. 5.1). Для вибору статистик необхідно відзначити відповідні відзначення. Діалогове вікно для редагування ідентифікаторів змінних викликається кнопкою "*Calculate Now*" (рис. 5.2). Обрані статистики із заданими ідентифікаторами змінних будуть обчислені й експортовані в середовище MATLAB після натискання кнопки "ОК" (рис. 5.1).

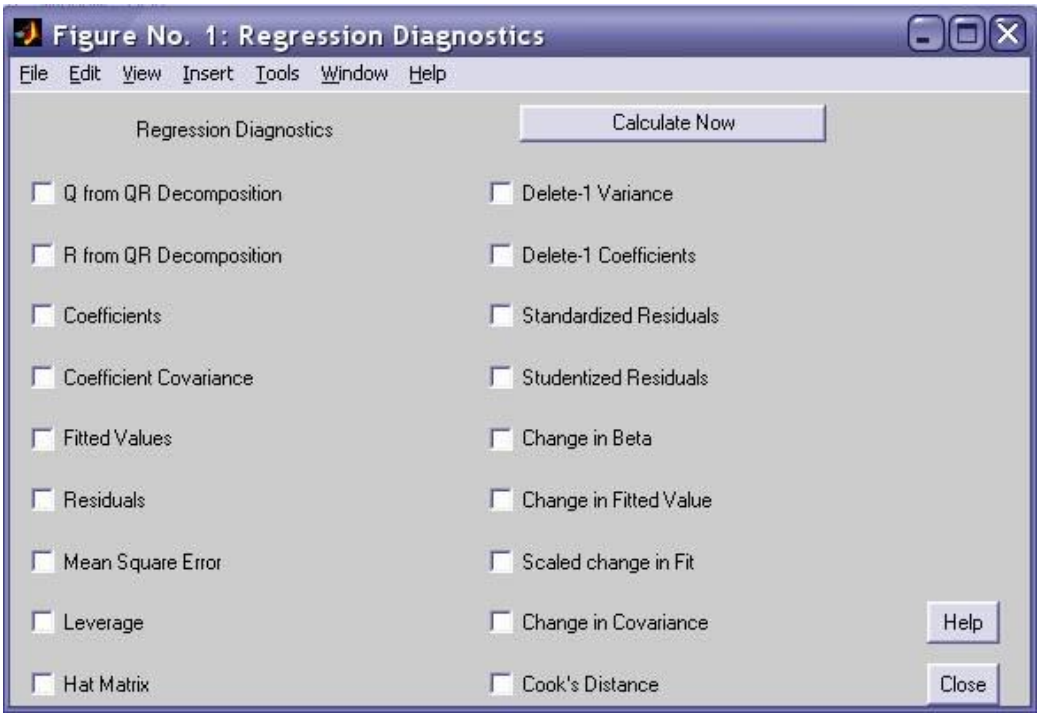


Рис. 5.1. **Графічне вікно вибору статистик множинної регресійної моделі**

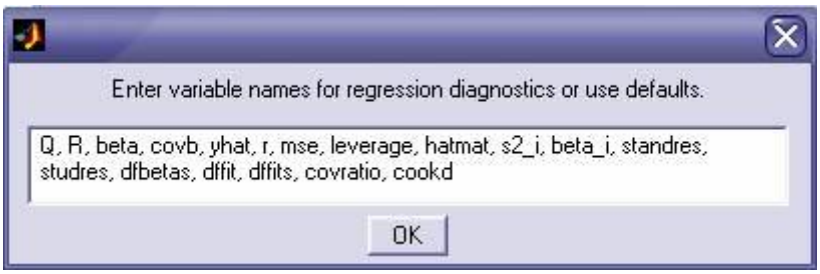


Рис. 5.2. **Графічне вікно ідентифікаторів статистик множинної регресійної моделі в середовищі *MATLAB***

Передбачено наступні види регресійних моделей (табл. 5.1).

Таблиця 5.1

Значення вхідного аргументу *'model'* функції *regstats*

Значення '*model*'	**Склад ефектів множинної регресійної моделі**
'linear'	Лінійна модель, що включає лінійні ефекти факторів і постійний член. Приймається за замовчуванням
'interaction'	Лінійна модель, що включає лінійні ефекти й ефекти взаємодії факторів, постійний член
'quadratic'	Квадратична модель, що включає квадратичні ефекти й ефекти взаємодії факторів
'purequadratic'	Квадратична модель, що включає квадратичні й лінійні ефекти факторів, постійний член

Послідовність коефіцієнтів множинної регресійної моделі відповідає їх порядку, використовуваному у функції `x2fx`.

Статистики множинної регресійної моделі, які наведені на рис. 5.1, описані в табл. 5.1. Послідовність статистик на рис. 5.1, 5.2 й у табл. 5.1 співпадає.

stats = regstats(responses,DATA,'model','whichstats') – функція повертає структуру *stats*, що містить статистики множинної регресійної моделі. Склад полів структури *stats* визначає вхідний аргумент *'whichstats'*. *'whichstats'* може бути заданий як константа у вигляді рядка, наприклад *'leverage'*, або масив ланцюгів утримуючі рядкові константи, наприклад {*'leverage' 'standres' 'studres'*}. Допустимі рядкові константи у формуванні *'whichstats'* наведені в табл. 5.1.

Таблиця 5.2

Значення вихідних параметрів функції *regstats*

Значення строкової константи	**Статистика, що повертається**
'Q'	Матриця *Q* у *QR* розкладання матриці незалежних змінних *DATA*
'R'	Матриця *R* у *QR* розкладання матриці незалежних змінних *DATA*
'beta'	Вектор точкових оцінок коефіцієнтів регресійної моделі
'covb'	Коваріційна матриця коефіцієнтів регресійної моделі
'yhat'	Вектор залежної змінної, обчислений за отриманою регресійною моделлю для вихідних значень незалежних змінних *DATA*
'r'	Вектор залишків
'mse'	Середня квадратична похибка
'leverage'	Вектор ступенів впливу окремих спостережень на коефіцієнти регресійної моделі
'hatmat'	Матриця, що проектує вектор спостережень залежної змінної на вектор спостережень залежної змінної, обчисленої за регресійною моделлю
's2_i'	Вектор середніх квадратичних похибок залежної зміної, отриманих без обліку поточного спостереження
'beta_i'	Матриця коефіцієнтів регресійної моделі. Стовпці матриці – вектори коефіцієнтів регресійної моделі, отримані послідовно без обліку одного зі спостережень
'standres'	Вектор стандартизованих залишків (залишків, ділених на величину їх середнього квадратичного відхилення)
'studres'	Вектор стандартизованих залишків без врахування поточного спостереження (залишків, поділених на величину відповідного елемента вектора *'s2_i'*)
'dfbetas'	Матриця впливу коефіцієнтів регресійної моделі. Стовпці матриці – вектори коефіцієнтів регресійної моделі, які не містять відповідного коефіцієнта, отримані послідовно без урахування одного зі спостережень
'dffit'	Вектор виправлень до значень залежної змінної, розрахованих без урахування поточного спостереження.
'dffits'	Вектор виправлень до значень залежної змінної, обчислених без урахування поточного спостереження – '*dffit*', нормований на величину стандартної похибки
'covratio'	Вектор відношень узагальненої дисперсії у визначенні значень коефіцієнтів регресійної моделі без поточного спостереження до узагальненої дисперсії коефіцієнтів регресії з урахуванням усіх спостережень
'cookd'	Вектор відстаней Кука. Елементи вектора відстаней Кука знаходять як нормалізоване виправлення до вектора коефіцієнтів без врахування поточного спостереження.
'all'	Створення структури *stats*, що включає всі перелічені статистики

Для отримання більш детальної інформації з кожної статистики використовують кнопку *Help* у вікні вибору розраховують статистик множинної регресійної моделі (рис. 5.1).

STEPWISE **Покрокова регресія в інтерактивному режимі**

Синтаксис

```
stepwise(X,y)
stepwise(X,y,inmodel)
stepwise(X,y,inmodel,alpha)
```

Опис

stepwise(*X,y*) – функція дозволяє отримати в інтерактивному режимі регресійну модель для залежної змінної *y* від незалежних змінних – стовпців матриці X. Залежна змінна *y* задається як вектор. Число елементів вектора *y* має дорівнювати кількості рядків матриці *X*. Функція відображує три графічних вікна для управління процесом покрокової регресії. Елементи управління в графічних вікнах призначені для видалення та додавання факторів, а також відображення статистик, що характеризують поточну регресійну модель.

Графік значень коефіцієнтів регресії й їх 95 % довірчих інтервалів дозволяє включати або видаляти фактори з регресійної моделі в інтреактивному режимі. Значення коефіцієнтів регресії та меж їх довірчих інтервалів, включені в модель, відображуються зеленими кольорами. Коефіцієнти регресії, виключені з регресійної моделі, виділяються червоними кольорами (рис. 5.3).

Включення або виключення фактору з регресійної моделі виконують клацанням лівої кнопки миші на відповідній лінії графіка. Межі довірчих інтервалів коефіцієнтів регресії, які значущо відрізняються від нуля, відображуються суцільними лініями. Межі довірчих інтервалів коефіцієнтів, що статистично не значущо відрізняються від нуля, відображуються штриховими лініями, які перетинають вертикальну нульову лінію. Значення коефіцієнта, не включеного в модель, обчислюють із припущення про його включення до складу поточної регресійної залежності.

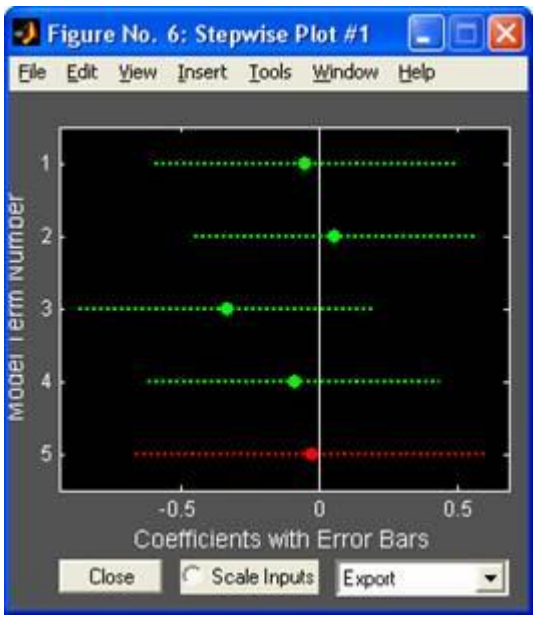


Рис. 5.3. **Діалогове вікно функції *Stepwise***

Кнопка *Scale Inputs* слугує для нормалізації центрованих значень елементів стовпців матриці незалежних змінних *X* на величину їх середнього квадратичного відхилення.

Кнопка *Close* дозволяє закрити графічні вікна.

Меню *Export* використовують для експорту результатів покрокової регресії на поточному кроці в робочу область *MATLAB* (рис. 5.4).

Рис. 5.4. **Приклад використання меню *Export***

Призначення пунктів меню *Export:*

Parameters – експорт вектора коефіцієнтів регресії;

Confidence Intervals – експорт матриці значень меж довірчих інтервалів коефіцієнтів регресії;

Terms In – експорт вектора номерів незалежних змінних, включених у регресійну модель;

Terms Out – експорт вектора номерів незалежних змінних, виключених з регресійної моделі;

All – експорт всіх зазначених вище параметрів.

Після вибору будь якого пункту меню *Export* буде відображене діалогове вікно, призначене для зміни ідентифікаторів експортованих змінних, заданих за замовчуванням. У виборі пукнта *All* будуть наведені наступні ідентифікатори експортованих змінних (рис. 5.5).

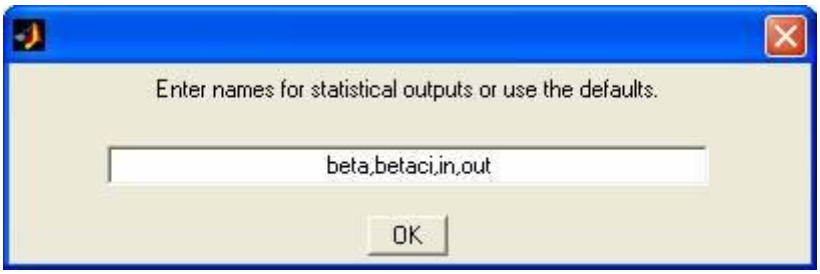


Рис. 5.5. **Ідентифікатори експортованих змінних**

Після натискання кнопки "OK" змінні із заданими ідентифікаторами будуть експортовані в середовище *MATLAB.*

Таблиця з параметрами покрокової регресії (рис. 5.6) містить у числовій формі інформацію, наведену на графіку значень коефіцієнтів. Таблиця містить наступні стовпці: *Colomn#* – номер стовпця матриці незалежних змінних, *Parameter* – значення коефіцієнта регресії; *Lower Confedence Intervals, Upper Confedence Intervals* – нижня та верхня межі довірчого інтервалу коефіцієнта регресійної моделі. Крім таблиці з параметрами регресійної моделі в графічному вікні наведені: *RMSE* – корінь квадратний із середньої квадратичної помилки, *R-square* – коефіцієнт детермінації, що показує, яка частина загальної дисперсії може бути пояснена регресійною моделлю, *F* – статистика Фішера, що відповідає регресійної моделі, *P* – рівень значущості статистики *F*.

Рис. 5.6. **Вихідні дані функції *Stepwise***

Рядки таблиці параметрів регресії, виділені зеленими кольорами, відповідають факторам, включеним у модель. Фактори, виділені червоними кольорами

виключені з регресійної моделі. Для включення або виключення факторів з регресійної моделі використовують аналогічну техніку.

У графічному вікні розв'язків відображується залежність значень коренів квадратних із середньої квадратичної похибки регресійних моделей та меж довірчих інтервалів *RMSE* від номера кроку. Кожне включення або видалення фактору з регресійної моделі у графічних вікнах призведе до додавання відповідного графіка у вікні *Figure No.* 5: *Stepwise History* #1 (рис. 5.7).

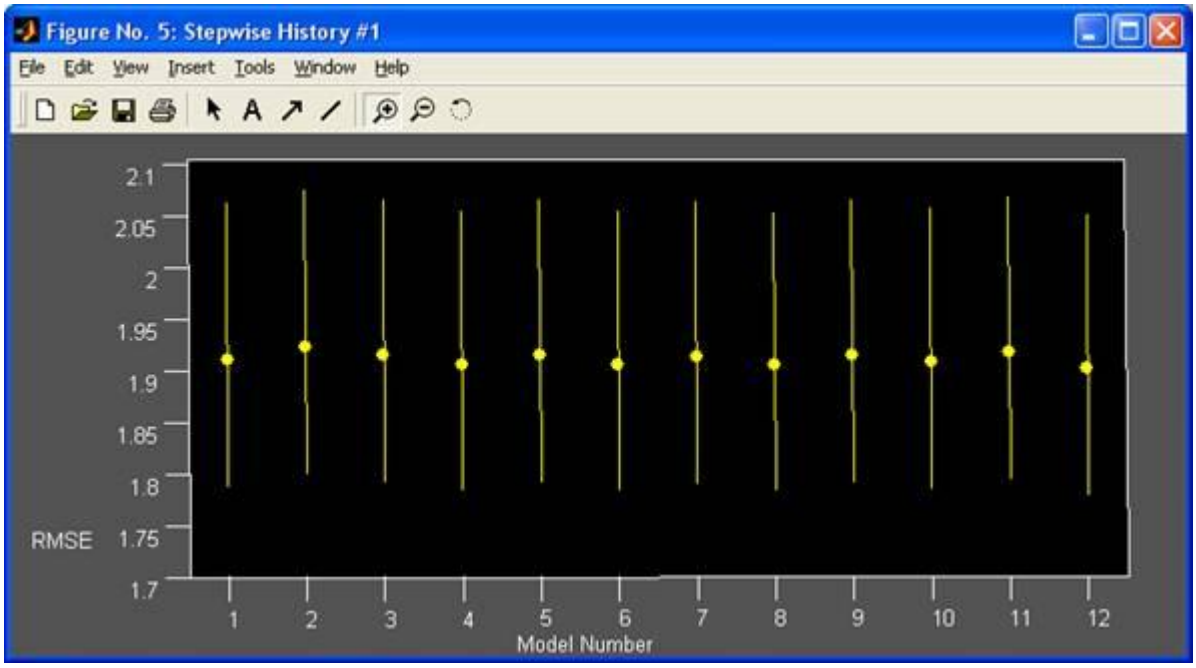


Рис. 5.7. **Графічне вікно функції *Stepwise***

Повернення до попередньої моделі виконують клацанням лівої кнопки миші на відповідному значенні *RMSE* у графічному вікні розв'язків. Нові параметри регресійної моделі відображуються в нових графічних вікнах, аналогічних наведеним на рис. 5.7.

stepwise(*X*,*y*,*inmodel*) – вхідний аргумент *inmodel* дозволяє управляти початковою множиною факторів, включених у регресійну модель. Елементи вектора *inmodel* є номерами факторів, тобто номерами стовпців матриці *X*.

stepwise(*X*,*y*,*inmodel*,*alpha*) – вхідний аргумент *alpha* задає рівень значущості для обчислення меж довірчого інтервалу коефіцієнтів регресійної моделі. *alpha* служить для перевірки гіпотези про статистичну значущість кожного фактора в регресійній моделі. За замовчуванням, $alpha = 1 - (-0.025)^{(1/p)}$, де p – кількість стовпців матриці *X*. Це значення *alpha*

відповідає 95 % довірчої ймовірності. Довірчий інтервал обчислюють для регресійної моделі за всім діапазоном зміни значень незалежних змінних (використовують метод Бонферроні).

RIDGE

Обчислення гребеневих оцінок параметрів лінійної регресійної моделі (Рідж-регресія)

Синтаксис

```
b = ridge(y,X,k)
```

Опис

`b = ridge(y,X,k)` – функція призначена для обчислення гребеневих оцінок параметрів лінійної регресійної моделі: $y = X\beta + \varepsilon$, де y – вектор значень залежної змінної; β – вектор коефіцієнтів лінійної моделі; X – матриця значень незалежних змінних; ε – вектор випадкових факторів, розподілених за нормальним законом з нульовим математичним сподіванням і σ^2 дисперсією $\varepsilon \sim N(0, \sigma^2 I)$.

Розмірності векторів значень залежної змінної y і збурюючих факторів ε - $n \times 1$, де n – кількість спостережень. Розмірність матриці X дорівнює $n \times p$, де p – кількість незалежних змінних. Стовпці матриці X відповідають незалежним змінним, рядки – спостереженням. Розмірність вектора коефіцієнтів лінійної регресійної моделі складе $p \times 1$.

Гребеневі оцінки параметрів лінійної регресійної моделі $\hat{\beta}$ обчислюються з наступного вираження:

$$b = \hat{\beta} = (X^T X + kI)^{-1} X^T y, \qquad (5.2)$$

де k – параметр рідж-регресії, задається як скалярна величина.

Якщо k = 0, то гребеневі оцінки параметрів лінійної регресійної моделі b збігаються із точковими оцінками, отриманими методом найменших квадратів. Якщо величина k збільшується, то постійне зміщення коефіцієнтів b збільшується, а дисперсія оцінки зменшується. Для матриці незалежних змінних X з низьким числом обумовленості зменшення величини дисперсії оцінок коефіцієнтів відбувається швидше, ніж компенсується величина постійного зміщення.

5.3. Розв'язування типової задачі в середовищі *MATLAB*

Задача. За даними двадцяти (n = 20) сільськогосподарських районів потрібно побудувати регресійну модель урожайності на основі показників:

y – урожайність зернових культур (ц/га);

x_1 – кількість колісних тракторів (наведеної потужності) на 100 га;

x_2 – кількість зернозбиральних комбайнів на 100 га;

x_3 – кількість знарядь поверхневої обробки ґрунту на 100 га;

x_4 – кількість добрив, що витрачають на гектар;

x_5 – кількість хімічних засобів для оздоровлення рослин, що витрачають на гектар.

Вихідні дані для аналізу наведені в табл. 5.3.

Для початку роботи необхідно створити новий М-файл. Для цього з меню *File* вибрати опцію *New*, а потім M-File. У вікні, що з'явилося, редактора М-файлів ввести вихідні дані у вигляді масивів-стовпців.

Таблиця 5.3

Вихідні дані для аналізу

№ п/п	y	x_1	x_2	x_3	x_4	x_5
1	9.7	1.59	0.26	2.05	0.32	0.14
2	8.4	0.34	0.28	0.46	0.59	0.66
3	9.0	2.53	0.31	2.46	0.30	0.31
4	9.9	4.63	0.40	6.44	0.43	0.59
5	9.6	2.16	0.26	2.16	0.39	0.16
6	8.6	2.16	0.30	2.69	0.32	0.17
7	12.5	0.68	0.29	0.73	0.42	0.23
8	7.6	0.35	0.26	0.42	0.21	0.08
9	6.9	0.52	0.24	0.49	0.20	0.08
10	3.5	3.42	0.31	3.02	1.37	0.73
11	9.7	1.78	0.30	3.19	0.73	0.17
12	10.7	2.40	0.32	3.30	0.25	0.14
13	12.1	9.36	0.40	11.51	0.39	0.38
14	9.7	1.72	0.28	2.26	0.82	0.17
15	7.0	0.59	0.29	0.60	0.13	0.35
16	7.2	0.28	0.26	0.30	0.09	0.15
17	8.2	1.64	0.09	1.44	0.20	0.08
18	8.4	0.09	0.22	0.05	0.43	0.20
19	13.1	0.08	0.25	0.03	0.73	0.20
20	8.7	1.36	0.26	1.17	0.99	0.42

Для того щоб скористатися функцією *regstats*, необхідно підготувати загальний масив даних, об'єднавши вихідні фактори.

Фрагмент М-файла

```
%% Лінійна множинна регресія: мультиколінеарність факторів
clc % Очистити командне вікно
clear all
%% Вихідні дані:
x1 = [1.59 0.34 2.53 4.63 2.16 2.16 0.68 0.35 0.52 3.42 1.78 2.40
9.36 1.72 0.59 0.28 1.64 0.09 0.08 1.36]';
x2 = [0.26 0.28 0.31 0.40 0.26 0.30 0.29 0.26 0.24 0.31 0.30 0.32
0.40 0.28 0.29 0.26 0.09 0.22 0.25 0.26]';
x3 = [2.05 0.46 2.46 6.44 2.16 2.69 0.73 0.42 0.49 3.02 3.19 3.30
11.51 2.26 0.60 0.30 1.44 0.05 0.03 1.17]';
x4 = [0.32 0.59 0.30 0.43 0.39 0.32 0.42 0.21 0.20 1.37 0.73 0.25
0.39 0.82 0.13 0.09 0.20 0.43 0.73 0.99]';
x5 = [0.14 0.66 0.31 0.59 0.16 0.17 0.23 0.08 0.08 0.73 0.17 0.14
0.38 0.17 0.35 0.15 0.08 0.20 0.20 0.42]';
y = [9.70 8.40 9.00 9.90 9.60 8.60 12.50 7.60 6.90 13.50 9.70 10.7
12.10 9.70 7.00 7.20 8.20 8.40 13.10 8.70]';
X = [x1 x2 x3 x4 x5];
```

З метою попереднього аналізу необхідно побудувати матрицю коефіцієнтів кореляції, використовуючи функцію *corrcoef*.

Фрагмент М-файла

```
% Матриця парних коефіцієнтів кореляції:
R = corrcoef([y X])
```

```
>> R =
1.0000    0.4303    0.3788    0.3995    0.5773    0.3321
0.4303    1.0000    0.6314    0.9841    0.1104    0.3410
0.3788    0.6314    1.0000    0.6797    0.1554    0.4911
0.3995    0.9841    0.6797    1.0000    0.0622    0.2962
0.5773    0.1104    0.1554    0.0622    1.0000    0.5706
0.3321    0.3410    0.4911    0.2962    0.5706    1.0000
```

Аналіз матриці парних коефіцієнтів кореляції показує, що результівна ознака y найбільш тісно пов'язана з показником x_4 – кількістю добрив, що витрачають на гектар ($r_{yx4} = 0{,}58$).

Отже, зв'язок між факторами досить тісний. Так, існує практично функціональний зв'язок між кількістю колісних тракторів (x_1) і кількістю знарядь поверхневої обробки ґрунту x_3 (r_{x1x3} = 0,98), що свідчить про наявність мультиколінеарності в моделі.

Спочатку слід побудувати рівняння множинної регресії, ігноруючи факт присутності мультиколінеарності. За викликом функції `regstats` відобразиться графічне вікно вибору статистик множинної регресії, у якому необхідно поставити позначки в такий спосіб (рис. 5.8).

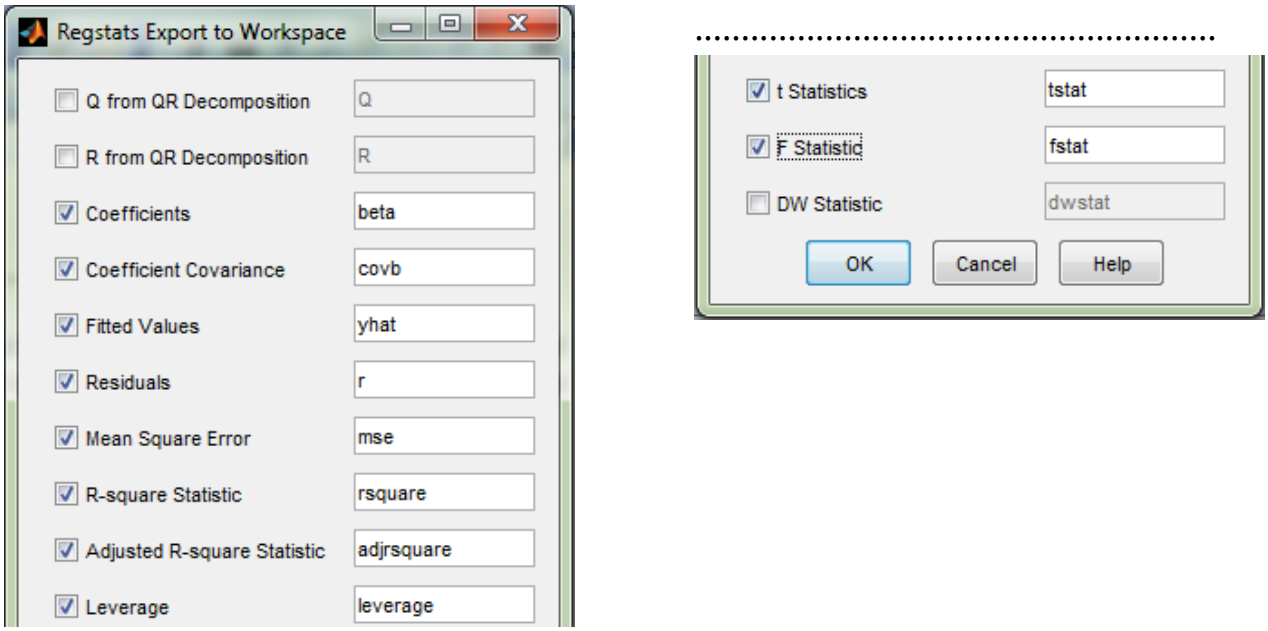


Рис. 5.8. **Діалогове вікно функції *regstats***

Функція `regstats` обчислює всі зазначені характеристики множинної регресійної моделі й експортує їх у середовище *MATLAB*. Інтерпретовані результати обчислення й експорту наступні:

```
>> Регресійне рівняння врожайності
Y = 5.4251 + 0.9348*x1 + 9.2079*x2 - 0.5482*x3 +
3.9175*x4 - 3.0562*x5

>> Рівняння адекватне

>> Значення t-статистик Стьюдента:
 0.8092
 0.9430
-0.5648
```

```
 2.7623
-1.0777

>> Коефіцієнт регресії b1 не є надійним
>> Коефіцієнт регресії b2 не є надійним
>> Коефіцієнт регресії b3 не є надійним
>> Коефіцієнт регресії b4 можна вважати надійним
>> Коефіцієнт регресії b5 не є надійним
```

Для обчислення регресійних статистик використовують експортні дані функції *regstats*.

Фрагмент М-файла

```
%% Регресійна статистика:
fprintf('Індекс множинної кореляції R')
R_mnog = sqrt(1 - sum(r.^2)/sum((y - mean(y)).^2))
fprintf('Коефіцієнт детермінації R-квадрат')
rsquare
fprintf('Нормований R-квадрат')
adjrsquare
```

```
>> Індекс множинної кореляції R

R_mnog =

    0.7190

>> Коефіцієнт детермінації R-квадрат

rsquare =

    0.5169

>> Нормований R-квадрат

adjrsquare =

    0.3444
```

Дисперсійний аналіз проводять за експортними даними функції *regstats*.

Фрагмент М-файла

```
%% Дисперсійний аналіз
fprintf('dfr - число ступеней свободи факторної дисперсії')
fstat.dfr
fprintf('dfe - число ступеней свободи залишкової дисперсії')
fstat.dfe
fprintf('SSr - факторна дисперсія')
fstat.ssr
fprintf('SSe - залишкова дисперсія')
fstat.sse
fprintf('F-критерій Фішера')
fstat.f
```

```
>> dfr - кількість ступеней свободи факторної дисперсії

     5

>> dfe - кількість ступеней свободи залишкової дисперсії

    14

>> SSr - факторна дисперсія

   38.3345

>> SSe - залишкова дисперсія

   35.8230

>> F-критерій Фішера

    2.9963
```

З рівняння видно, що статистично значущим є коефіцієнт регресії тільки з x_4, тому що $|t_4| = 2{,}7623 > t_{\text{расч}} = 2{,}1$ 448. Не піддаються економічній інтерпретації від'ємні значення коефіцієнтів регресії з x_3 і x_5, зі значення яких витікає, що підвищення кількості знарядь поверхневої обробки ґрунту (x_3) і засобів оздоровлення рослин (x_5) від'ємно позначається на врожайності. Таким чином, отримане рівняння регресії неприйнятне. Причиною цього може бути наявність мультиколінеарності. Для її усунення використовують ряд вбудованих функцій у *MATLAB*. Доцільно розглянути роботу двох з них: функцію *stepwise* і функцію *ridge*.

Для цього слід реалізувати алгоритм покрокового регресійного аналізу з виключенням змінних (функція *stepwise*), зважаючи на те, що в рівняння повинна ввійти тільки одна з тісно зв'язаних змінних (рис. 5.9 – 5.12).

Крок 1. Move X4 in

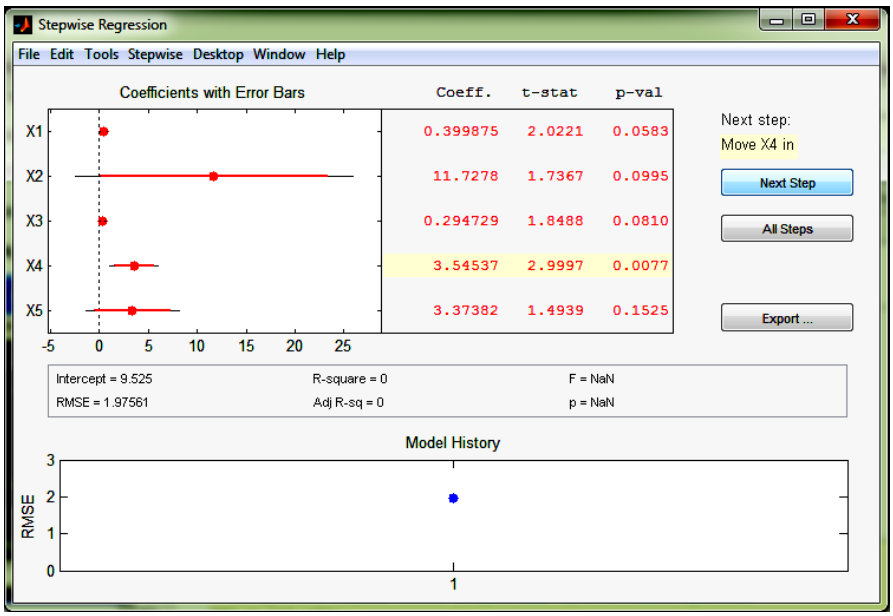


Рис. 5.9. **Включити в модель фактор X4**

Крок 2. Move X1 in

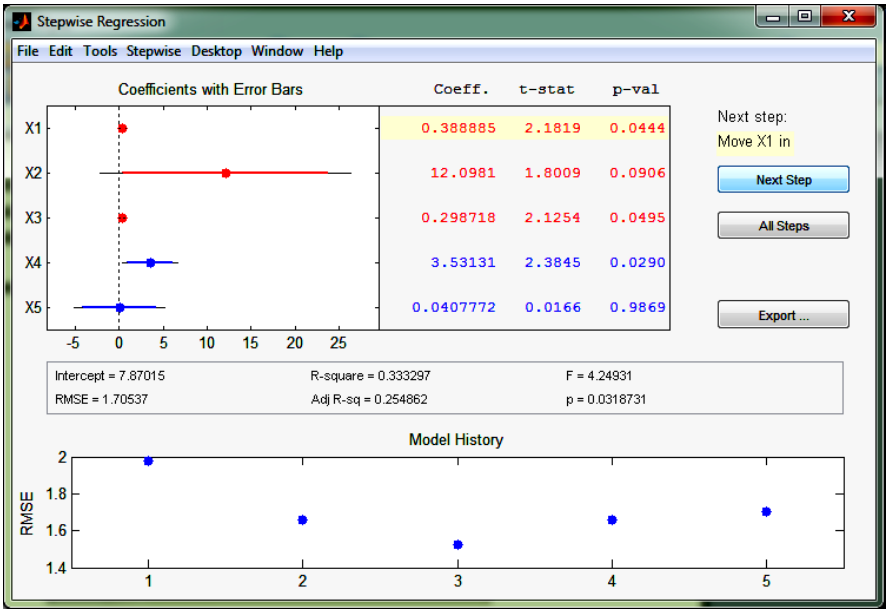


Рис. 5.10. **Включити в модель фактор X1**

Крок 3. Move X5 out

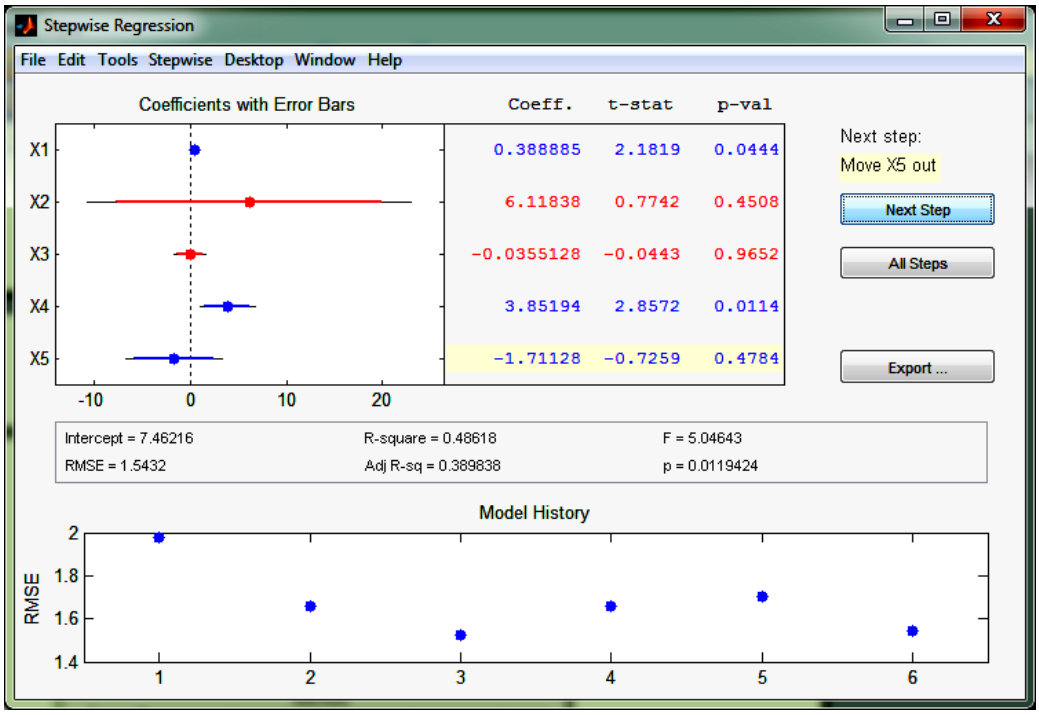


Рис. 5.11. **Виключити з моделі фактор X5**

Крок 4. Export

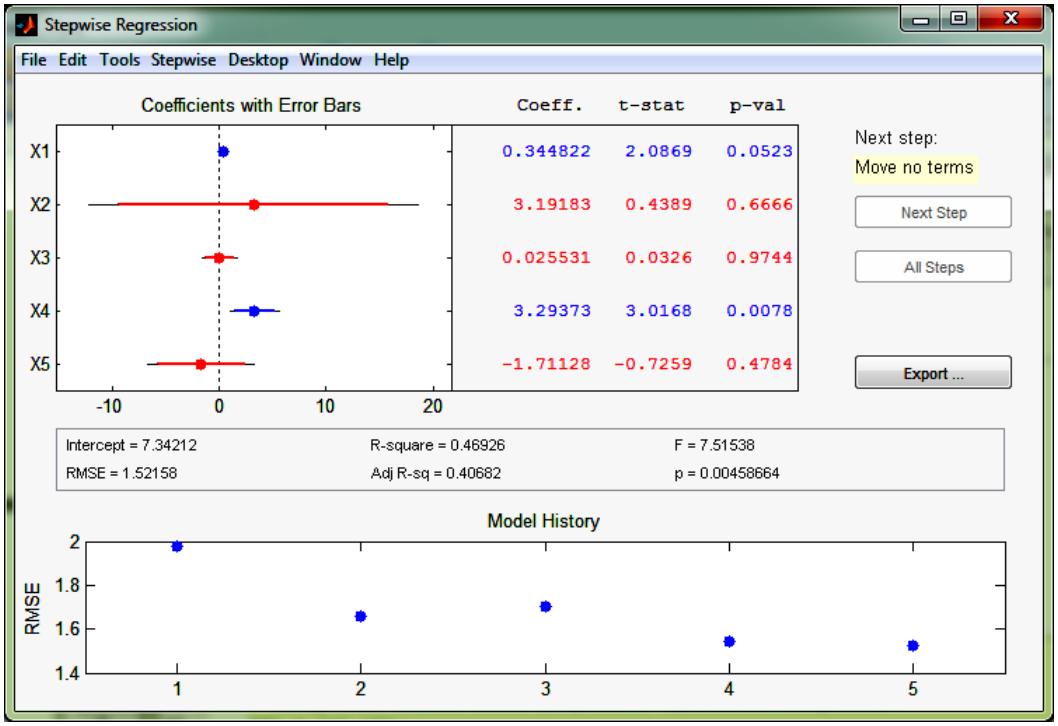


Рис. 5.12. **Експортувати дані у середовище *MATLAB***

Результати експорту моделі:

```
>> R =
    1.0000    0.4303    0.5773
    0.4303    1.0000    0.1104
    0.5773    0.1104    1.0000

>> Регресійне рівняння врожайності

    y_p = 7.3421 + 0.3448 *x1 + 3.2937 *x4

>> Рівняння адекватне
>> Значення t-статистик Стьюдента:

0.8092
0.9430

>> Коефіцієнт регресії b1 не є надійним
>> Коефіцієнт регресії b4 можна вважати надійним

>> Індекс множинної кореляції R

R_mnog =

    0.6850

>> Коефіцієнт детермінації R-квадрат

rsquare =

    0.5169

>> Нормований R-квадрат

adjrsquare =

    0.3444
```

```
>> dfr - кількість ступеней свободи факторної дисперсії

     2

>> dfe - кількість ступеней свободи залишкової дисперсії

    17

>> SSr - факторна дисперсія

   34.7992

>> SSe - залишкова дисперсія

   39.3583

>> F-критерій Фішера

    7.5154
```

З рівняння регресії видно, що збільшення на одиницю кількості тракторів на 100 га ріллі призводить до зростання врожайності зернових у середньому на 0,345 ц/га (b_1 = 0,3 448).

Для порівняльного аналізу методів усунення мультиколінеарності використовують інший спосіб – рідж-регресію (функція *ridge* в *MATLAB*).

Фрагмент М-файла

```
%% 2 спосіб Рідж-регресія:
k = [0.0:0.01:1];
n = size(k')
b=zeros(n(1),6)
     for i=1:n(1)
          b(i,:) = ridge(y,[ones(20,1) X], k(i));
     end
% Графічне відображення отриманих результатів
plot(k',b(:,1),k',b(:,2),k',b(:,3),k',b(:,4),k',b(:,5),k',b(:,6))
title('Залежність бета-коефіцієнтів від параметра k')
legend('b0','b1','b2','b3','b4','b5',-1)
xlabel('k'),ylabel('b','FontName','Symbol')
grid on
```

На рис. 5.13 наведені графіки гребеневого сліду для всіх шести параметрів множинної регресії.

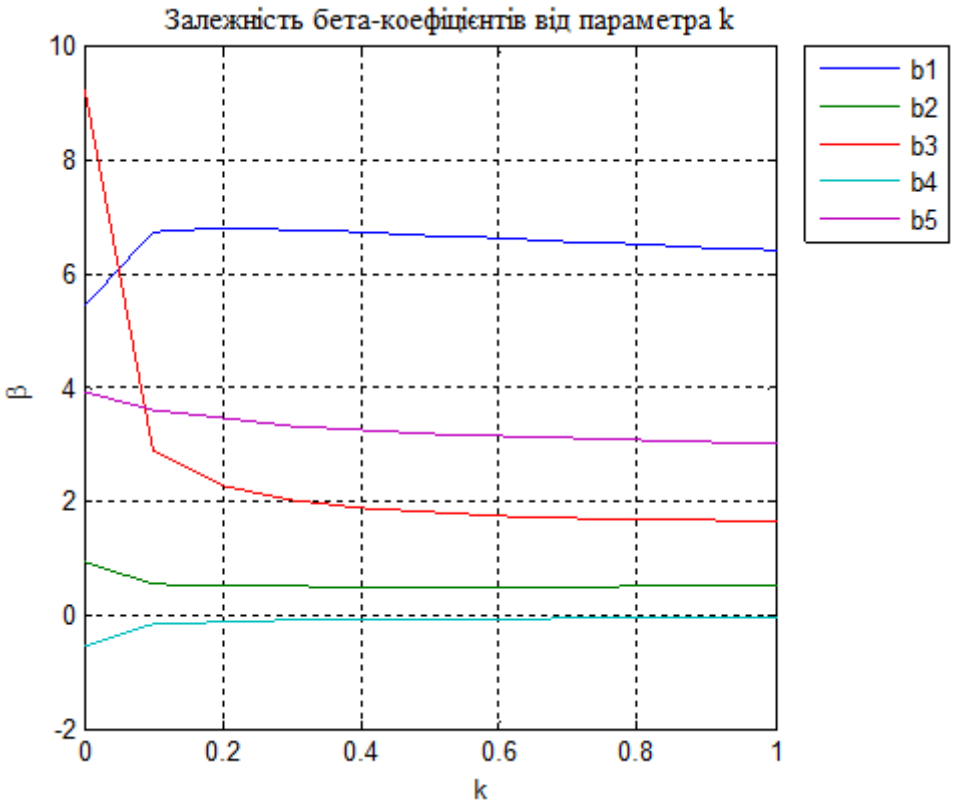


Рис. 5.13. **Залежність β-коефіцієнтів від параметра *k***

У табл. 5.4 наведені значення параметрів "чистої регресії" залежно від параметра *k*.

Таблиця 5.4

Залежність параметрів моделі від параметра *k*

k	**0**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1**
b_0	5,42	6,72	6,78	6,75	6,71	6,66	6,60	6,55	6,50	6,45	6,40
b_1	0,93	0,55	0,50	0,49	0,48	0,48	0,48	0,48	0,49	0,49	0,50
b_2	9,20	2,90	2,24	2,01	1,88	1,80	1,75	1,71	1,67	1,65	1,62
b_3	-0,54	-0,17	-0,12	-0,10	-0,08	-0,07	-0,07	-0,06	-0,05	-0,05	-0,04
b_4	3,91	3,60	3,44	3,33	3,24	3,18	3,13	3,09	3,05	3,02	2,99
b_5	-3,05	-1,41	-0,87	-0,52	-0,27	-0,08	0,05	0,17	0,27	0,35	0,42

Обчислення коефіцієнтів *R*2 і *RR* легко здійснити, використовуючи наступний цикл.

Фрагмент М-файла

```
% Розрахунок характеристик моделі залежно від параметра k:
b = zeros(size(X'*X,1),size(k,2));
yfit = zeros(size(y,1),size(k,2));
```

```
R2=zeros(size(k,2),1);
RR=zeros(size(k,2),1);
for i=1:size(k,2)
b(:,i) = inv(X'*X + k(i)*eye(size(X,2)))*X'*y;%ридж-регресія
yfit(:,i) = b(1,i) + b(2,i)*x1 + b(3,i)*x2 + b(4,i)*x3 + b(5,i)*x4
+ b(6,i)*x5;% розрахункові значення y
end
for i=1:size(k,2)
     R2(i,1) = 1 - sum((y - yfit(:,i)).^2)/sum((y - mean(y)).^2);%
R2
     RR(i,1)  =  sum((yfit(:,i)  -  mean(yfit(:,i))).^2)/sum((y  -
mean(y)).^2);% RR
end
% Графічне відображення отриманих результатів
plot(k,R2,k,RR)
legend('R2','RR',-1)
grid on
```

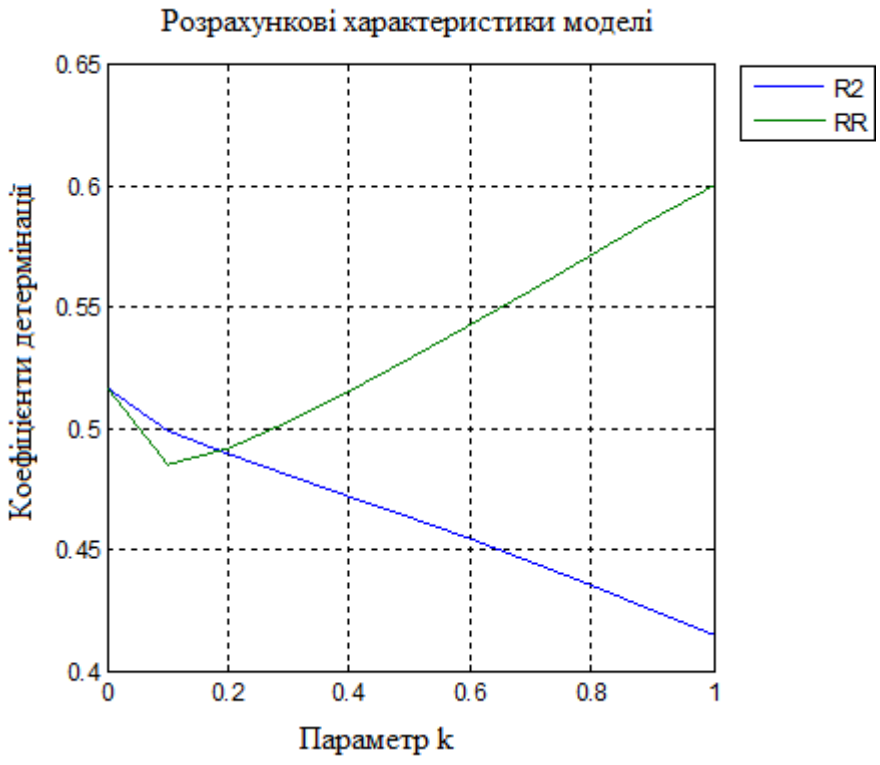


Рис. 5.14. **Залежність коефіцієнтів *R2* та *RR* від параметра *k***

Можна побачити, що коефіцієнт детермінації зменшується зі зростанням рідж-параметра *k*.

Рівняння "чистої" регресії:

```
>> Регресійне рівняння врожайності

y_p = 6.6 + 0.48 *x1 + 1.75 *x2 – 0.07 *x3 + 3.13 *x4 +
0.05 *x5

>> Рівняння адекватне
```

На підставі проведених обчислень можна зробити висновок, що із двох моделей множинної регресії (покрокова регресія – модель1, рідж-регресія – модель 2) перевагу необхідно віддати першій.

Коефіцієнт детермінації 0,5159 свідчить про те, що тільки 51,59 % варіації врожайності пояснюється показниками (x_1 і x_4), тобто забезпеченістю рослинництва тракторами та добривами. Інша частина варіації обумовлена дією неврахованих факторів (x_2, x_3, x_5, погодними умовами та ін.).

Запитання для самоперевірки

1. Наведіть передумови МНК. Які наслідки їх виконання або невиконання?
2. Які властивості повинні мати оцінки, отримані за МНК?
3. Поясніть сутність проблеми мультиколінеарності у побудові множинної регресійної моделі.
4. Які основні наслідки мультиколінеарності?
5. Як можна виявити мультиколінеарність?
6. Наведіть основні методи усунення мультиколінеарності. Поясніть їх відмінність.

Завдання для лабораторної роботи

Задача. Є дані про споживання деякого продукту Y (у. о.) залежно від рівня урбанізації (частки міського населення) $X1$, відносного освітнього рівня $X2$ і відносних заробітків $X3$ для дев'яти географічних районів: (p_1 – кількість букв у повному імені, p_2 – кількість букв у прізвищі) (табл. 5.5).

Таблиця 5.5

Вихідні дані задачі

***X*1**	***X*2**	***X*3**	***Y***
1	2	3	4
167,1 + p_1	42,2 – p_2	11,2	31,9
48,6 + p_1	10,6 – p_2	13,2	174,4

Закінчення табл. 5.5

1	2	3	4
42,6 + p_1	10,6 – p_2	28,7	160,8
39,0 + p_1	10,4 – p_2	26,1	162,0
34,7 + p_1	9,3 – p_2	30,1	140,8
44,5 + p_1	10,8 – p_2	8,5	174,6
39,1 + p_1	10,7 – p_2	24,3	163,7
40,1 + p_1	10,0 – p_2	18,6	174,5
45,9 + p_1	12,0 – p_2	20,4	185,7

За даними, що містяться в табл. 5.5 виконати наступні дії.

1. Побудувати модель множинної регресії, обчислити кореляційно-регресійний аналіз моделі засобами *Statistic Toolbox*.

2. Визначити наявність або відсутність мультиколінеарності факторів.

3. За наявності мультіколінеарності усунути її за допомогою вбудованих функцій MATLAB.

4. Побудувати модель множинної регресії, використовуючи тільки значущі фактори.

5. Провести порівняльний аналіз моделей і дати економічну інтерпретацію отриманим результатам.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

Лабораторна робота 6
Гетероскедастичність і методи її визначення

Мета роботи: навчитися оцінювати наявність ефекту гетероскедастичності, вивчити тести її визначення, а також освоїти застосування функції `robustfit` середовища *MATLAB*, що реалізує зважений метод найменших квадратів.

Основні задачі лабораторної роботи:

1. Побудувати модель регресії на підставі статистичних даних.

2. Побудувати графічне подання відхилень. На підставі графічних даних висунути гіпотезу про наявність гетероскедастичності.

3. Провести аналітичну перевірку гетероскедастичності, використовуючи тест рангової кореляції Спірмена та тест Голдфелда – Квандта.

4. Оцінити параметри регресійної моделі в умовах гетероскедастичності.

5. Скориставшись зваженим методом найменших квадратів (WLS), зменшити гетероскедастичність.

6. Провести порівняльний аналіз отриманих моделей.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

6.1. Сутність гетероскедастичності

Однією з ключових передумов МНК є умова сталості дисперсій випадкових відхилень, тобто $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$ для будь-яких спостережень i й j. Виконання даної передумови називається гомоскедастичністю, невиконання − гетероскедастичністю. У ряді невиконання даної передумови, тобто за гетероскедастичності, наслідки застосування МНК будуть наступними.

1. Оцінки коефіцієнтів, як і раніше, залишаться незміщеними та лінійними.

2. Оцінки не будуть ефективними та навіть асимптотично ефективними. Збільшення дисперсії оцінок знижує ймовірність отримання максимально точних оцінок.

3. Дисперсії оцінок розраховуватимуться зі зміщенням. Зміщеність з'являється внаслідок того, що не пояснена рівнянням регресії дисперсія:

$$S^2 = \frac{\sum \varepsilon_i^2}{n - m - 1}, \tag{6.1}$$

яка використовується у процесі обчислення оцінок дисперсій всіх коефіцієнтів, не є незміщеною.

4. Тому всі висновки, отримані на основі відповідних *t*- і *F*-статистик, а також інтервальні оцінки будуть ненадійними. Отже, статистичні висновки, отримані у процесі стандартних перевірок якості оцінок, можуть бути помилковими та призводити до неправильних висновків щодо побудованої моделі.

На сьогодні не існує однозначного методу визначення гетероскедастичності, тому виявити її в кожному випадку досить складно. Проте для такої перевірки розроблено достатню кількість тестів і критеріїв для них. Найбільш популярними та наочними є: графічний аналіз відхилень, тест рангової кореляції Спірмена, тест Парка, тест Глейзера, тест Голдфельда − Квандта.

Графічний аналіз залишків. Даний метод є найпростішим, і візуальне подання відхилень дозволяє визначитися з наявністю гетероскедастичності. У цьому випадку на осі абсцис відкладаються значення (x_i) пояснювальної змінної

X (або лінійної комбінації пояснювальних змінних $Y = a + b_1 x_1 + ... + b_m x_m$), а на осі ординат відхилення ε_i або їх квадрати ε_i^2.

Тест рангової кореляції Спірмена. Використання даного тесту передбачає, що дисперсія відхилень буде збільшуватися або зменшуватися із збільшенням значень x. Тому для регресії, побудованої за допомогою МНК, абсолютні величини відхилень ε_i і значення x_i випадкової величини X будуть корельовані. Значення x_i і ε_i ранжують (упорядковуються за величинами). Потім визначають коефіцієнт рангової кореляції:

$$r_{x\varepsilon} = 1 - 6 \cdot \frac{\sum d_i^2}{n(n^2 - 1)}, \qquad (6.2)$$

де d_i – різниця між рангами x_i і ε_i.

Доведено, що якщо коефіцієнт кореляції $\rho_{x\varepsilon}$ для генеральної сукупності дорівнює нулю, то статистика $t = \frac{r_{x\varepsilon}\sqrt{n-2}}{\sqrt{1 - r_{x\varepsilon}^2}}$ має розподіл Стьюдента з кількістю ступенів свободи $v = n - 2$. Отже, якщо значення t-статистики перевищує $t_{\text{кр}} = t_{\frac{\alpha}{2}, n-2}$, то необхідно відхилити гіпотезу про дорівненість нулю коефіцієнта кореляції $\rho_{x,\varepsilon}$, а отже, і про відсутність гетероскедастичності. В іншому випадку гіпотеза про відсутність гетероскедастичності приймається.

Якщо в моделі регресії більше одної пояснювальної змінної то перевірка гіпотези може здійснюватися за допомогою t-статистики для кожної з них окремо.

Тест Парка. Р. Парк запропонував критерій визначення гетероскедастичності, який доповнює графічний метод деякими формальними залежностями. Передбачається, що дисперсія $\sigma_i^2 = \sigma^2(x_i)$ є функцією i-го значення x_i пояснювальної змінної. Парк запропонував таку залежність: $\sigma_i^2 = \sigma^2 x_i^{\beta} \varepsilon^{v_i}$. Після логарифмування даного виразу отримано $\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln x_i + v_i$. Оскільки дисперсії σ_i^2 зазвичай невідомі, тому їх замінюють оцінками ε_i^2.

Критерій Парка передбачає наступні етапи:

1) будують рівняння регресії $y_i = a + b_1 x_i + \varepsilon_i$;

2) для кожного спостереження визначають $\ln \varepsilon_i^2 = \ln(y_i - \hat{y}_i)^2$;

3) будують регресію $\ln\varepsilon_i^2 = \alpha + \beta \ln x_i + v_i$, де $\alpha = \ln\sigma^2$. Для множинної регресії будують цю залежність для кожної пояснювальної змінної;

4) перевіряють статистичну значущість коефіцієнта β на основі t-статистики $t = \dfrac{\beta}{\sigma_\beta}$. Якщо коефіцієнт β статистично значущий, то це означає наявність зв'язку між $\ln\varepsilon_i^2$ і $\ln x_i$, тобто гетероскедастичності в статистичних даних.

Застосування тесту Парка в деяких випадках може призвести до необґрунтованих висновків, тому він доповнюється іншими тестами.

Тест Глейзера. Вважається, що тест Глейзера за своєю сутністю аналогічний тесту Парка та доповнює його аналізом інших залежностей між дисперсіями відхилень σ_i та значеннями змінної x_i. Тут оцінюється регресійна залежність модулів відхилень $|\varepsilon_i|$ від $x|\varepsilon_i|$. У цьому випадку залежність моделюють таким рівнянням:

$$|\varepsilon_i| = \alpha + \beta x_i^k + v_i. \qquad (6.3)$$

Змінюючи значення k, можна побудувати різні регресії. Статистична значущість коефіцієнтів β в кожному випадку означає наявність гетероскедастичності. Якщо для декількох регресій коефіцієнт β виявиться значущим, то орієнтуються на найкращу з них. Як і в тесті Парка, в тесті Глейзера для відхилень v_i може порушуватись умова гетероскедастичності, проте в багатьох випадках запропоновані моделі є придатними для визначення гетероскедастичності.

Тест Голдфельда – Квандта. Тест Голдфельда – Квандта передбачає, що стандартне відхилення σ_ε пропорційне значенню змінної x у цьому спостереженні: $\sigma_{\varepsilon_i}^2 = \sigma^2 x_i^2$ ($= \overline{1, n}$. Також передбачається, що ε_i має нормальний розподіл, відсутня автокореляція і всі n спостережень упорядковуються за величиною x. Цю упорядковану вибірку розділяють на три приблизно рівні частини $k, n-2k, k$, відповідно. Для кожної з вибірок обсягу k оцінють її рівняння регресії та знаходять суми квадратів відхилень $S_1 = \sum_{i=1}^{k} \varepsilon_i^2$ і $S_3 = \sum_{i=n-k+1}^{k} \varepsilon_i^2$, відповідно.

За заданою довірчою ймовірністю $p, \alpha = 1 - p$ за F-таблицями знаходять межеву точку $F_{\alpha, k-m-1, k-m-1}$, де $m-$число факторів моделі; розраховують значення $F = \frac{S_3}{S_1}$. Якщо $F < F_{\alpha, k-m-1, k-m-1}$, то на рівні значущості α приймається гіпотеза про відсутність гетероскедастичності. В іншому випадку гіпотеза про відсутність гетероскедастичності відхиляється. Для множинної регресії тест зазвичай проводиться для того фактора, який максимально пов'язаний з σ_{ε_i}. Для цього обирають $k > m+1$. Для переконливості даний тест можна розрахувати для кожного фактора.

Метод зважених найменших квадратів (МЗНК). Якщо відомі σ_ε^2, то рекомендується застосовувати метод зважених найменших квадратів (МЗНК). Для цього слід розділити кожне спостереження на відповідне йому значення дисперсії. Наприклад, $y_i = a + b_1 x_i + \varepsilon_i$. Слід поділити обидві частини рівняння на відоме $\sigma_{\varepsilon_i} = \sqrt{\sigma_{\varepsilon_i}^2}$:

$$\frac{y_i}{\sigma_{\varepsilon_i}} = \frac{a}{\sigma_{\varepsilon_i}} + b_1 \frac{x_i}{\sigma_{\varepsilon_i}} + \frac{\varepsilon_i}{\sigma_{\varepsilon_i}}. \tag{6.4}$$

Нехай $\frac{y_i}{\sigma_{\varepsilon_i}} = y_i^*$, $\frac{x_i}{\sigma_{\varepsilon_i}} = x_i^*$, $\frac{\varepsilon_i}{\sigma_{\varepsilon_i}} = v_i$, $\frac{1}{\sigma_{\varepsilon_i}} = z_i$. Утворюється рівняння регресії без вільного члена, але з додатковою змінною, що пояснює Z, і з перетвореним відхиленням v_i:

$$y_i^* = az_i + b_1 x_i^* + v_i. \tag{6.5}$$

У цьому випадку для всіх v_i виконується умова гомоскедастичності.

Отже, метод зважених найменших квадратів містить такі етапи. *Перший етап* передбачає, що всі спостереження (x_i, y_i) ділять на відому величину σ_{ε_i}. Таким чином, спостереження з найменшими дисперсіями набувають найбільшої ваги. Отже, вони будуть значущими у процесі оцінювання коефіцієнтів регресії. Спостереження з максимальними дисперсіями, навпаки, набувають найменшої ваги та будуть менш значущими. На *другому етапі* за методом найменших квадратів для перетворених значень $\left(\frac{1}{\sigma_i}, \frac{x_i}{\sigma_i}, \frac{y_i}{\sigma_i}\right)$ будують рівняння регресії без вільного члена з гарантованими якостями оцінок.

На практиці рекомендують застосовувати декілька методів визначення гетероскедастичності та способів її коректування.

6.2. Теоретичні відомості про функції *MATLAB*, які використовуються в даній лабораторній роботі

ROBUSTFIT **Робастна регресія**

Синтаксис

```
b = robustfit(X,Y)
[b,stats] = robustfit(X,Y)
[b,stats] = robustfit(X,Y,'wfun',tune,'const')
```

Опис

b = robustfit(X,Y) – функція дозволяє отримати робастні оцінки параметрів регресійної моделі *b* для матриці незалежних змінних *X* і вектора значень залежної змінної. Стовпці матриці *X* є окремі незалежні змінні, рядки-спостереження. Кількість елементів вектора *Y* і рядків матриці *X* повинні дорівнювати. У *robustfit* реалізується ітераційний зважений метод найменших квадратів. Ваги на поточній ітерації обчислюють за допомогою біквадратичної функції від вектора залишків, обчислених на попередній ітерації. Використання такого алгоритму дозволяє задати менші значення ваги для спостережень, що мають більші відхилення від регресійної моделі відносно інших. Результати обчислення *b* менш чутливі до випадкових викидів у вибірці, ніж у використанні методу найменших квадратів.

[b,stats] = robustfit(X,Y) – функція повертає вектор точкових оцінок коефіцієнтів регресійної моделі *b* і структуру *stats*, що містить наступні поля:

stats.ols_s – корінь квадратний із середньої квадратичной похибки в обчисленні параметрів регресійної моделі методом найменших квадратів;

stats.robust_s – робастна оцінка стандартної похибки;

stats.mad_s – оцінка стандартної похибки на основі абсолютних відхилень залишків щодо їх медіани. Ця оцінка

використовується для нормування вектора залишків в ітеративній процедурі обчислення коефіцієнтів регресійної моделі;

stats.s – кінцева оцінка стандартної похибки. stats.s перевищує значення robust_s і обчислюється як зважене середнє від ols_s і robust_s;

stats.se – стандартна похибка точкових оцінок коефіцієнтів регресійної моделі;

stats.t – відношення вектора коефіцієнтів регресійної моделі b до їх стандартних похибок stats.se;

stats.p – рівень значущості для статистики stats.t;

stats.coeffcorr – оцінка коефіцієнтів кореляції оцінок коефіцієнтів регресійної моделі;

stats.w – вектор ваги у проведенні робастної регресії;

stats.h – вектор ступенів впливу спостережень на параметри регресійної моделі;

stats.dfe - кількість ступенів свободи стандартної похибки;

stats.R - матриця R з QR розкладання матриці незалежних змінних X.

У процесі робастної регресії розраховують коваріаційну матрицю коефіцієнтів рівняння регресії *V*. Обчислення *V* виконують як *V=inv(X'*X)*stats.s*^2. Стандартні похибки та коефіцієнти кореляції параметрів регресійної моделі використовують *V*.

[*b,stats*] = *robustfit*(*X,Y,'wfun',tune,'const'*) додаткові вхідні аргументи *'wfun', 'tune', 'const'* дозволяють задати функцію ваги, що узгоджує сталу, й вказати наявність або відсутність сталого члена. Передбачено наступні функції ваги (табл. 6.1).

Таблиця 6.1

Значення ваги функції *robustfit*

Вагова функція	Вид вагової функції	Узгоджувальна стала
1	2	3
'andrews'	*w* = (*abs*(r)<pi) .* sin(r) ./ r	1. 339
'bisquare'	*w* = (abs(r)<1) .* (1 - r.^2).^2	4. 685

Закінчення табл. 6.1

1	2	3
'cauchy'	`w = 1 ./ (1 + r.^2)`	2. 385
'fair'	`w = 1 ./ (1 + abs(r))`	1. 400
'huber'	`w = 1 ./ max(1, abs(r))`	1. 345
'logistic'	`w = tanh(r) ./ r`	1. 205
'talwar'	`w = 1 * (abs(r)<1)`	2. 795
'welsch'	`w = exp(-(r.^2))`	2. 985

Значення вхідної величини *r* вагової функції визначають з виразу *r=resid*/(*tune*s*sqrt*(1-*h*)), де: *resid* – вектор залишків від попередньої ітерації; *tune* – узгоджувальна стала; *h* – вектор ступенів впливу спостережень після обчислення параметрів рівняння регресії методом найменших квадратів; *s* – оцінка стандартного відхилення погрішності коефіцієнта рівняння регресії:

$$s = MAD/0.6745. \qquad (6.6)$$

Величина MAD є медіаною абсолютних відхилень залишків відносно їх медіани. Стала 0,6 745 призначена для отримання незміщеної оцінки *s* для нормального розподілу. Якщо в матриці незалежних змінних *X* задано *p* стовпців, включаючи сталий член, найменше ніж *p*-1 абсолютне відхилення виключається при обчисленні їх медіани.

У доповненні до наведеного списку вагових функцій вхідний параметр *'wfun'* може дорівнювати *'ols'*, що призведе до використання звичайного, не зваженого, методу найменших квадратів.

Вхідний параметр *tune* дозволяє задати нове значення узгоджувальної сталої, відмінної від наведених у табл. 6.1. Зменшення величини узгоджувальної сталої призведе до зменшення ваги більших за абсолютною величиною залишків і навпаки. Значення узгоджувальних сталих за замовчуванням, наведених у табл. 6.1, дозволяють отримати оцінки коефіцієнтів рівняння регресії, які приблизно в 95 % випадків також ефективні, як і оцінки, отримані методом найменших квадратів, за умови, що залежна змінна розподілена за нормальним законом без викидів.

Вхідний параметр *'const'* дозволяє додати вільний член у рівняння регресії, якщо *'const'='on'*, і видалити його, якщо *'const'='off'*. Значення за замовчуванням *'const'='on'*. Якщо необхідно отримати оцінку

вільного члена рівняння регресії, то необхідно вказати *'const'='on'*, а не використовувати одиничний стовпець у матриці *X*.

Функція користувача може бути використана як вагова функція. Вагова функція користувача може бути задана як функція-*inline*-функція або функція-*m*-функція. У цьому випадку як параметр *'wfun'* передається покажчик на функцію користувача, наприклад *@myfun*. Функція, обумовлена користувачем як вагарня, повинна отримати як вхідний аргумент вектор нормованих залишків. Його треба обчислити у вектор ваги та повернути останній як вихідний параметр.

FINV Обернена функція розподілу ймовірностей закону Фішера

Синтаксис

```
X = finv(P,V1,V2)
```

Опис

finv(P,V1,V2) слугує для обчислення значення квантилей закону Фішера для параметрів розподілу *V1*, *V2* і значення ймовірності *P*. Розмірність векторів або матриць *P*, *V1* й *V2* повинна бути однаковою. Розмірність скалярного параметра збільшується до розміру інших вхідних аргументів. Параметри *V1* і *V2* повинні бути додатними цілими числами. Значення ймовірності *P* повинне міститися в інтервалі [0 1]. Результат обчислення *X* – квантиль розподілу закону Фішера, що відповідає попаданню випадкової величини в інтервал (-∞ *X*] з імовірністю *P* з заданими значеннями кількості ступенів свободи *V1*, *V2*.

TINV Обернена функція розподілу ймовірностей закону Стьюдента

Синтаксис

```
X = tinv(P,V)
```

Опис

tinv(P,V) слугує для обчислення значень квантилей закону Стьюдента для значень імовірності *P* і ступеня свободи *V*. Розмірність векторів або матриць *P*, *V* повинна бути однаковою. Розмірність скалярного параметра збільшується до розмірності іншого вхідного аргументу. Значення числа

ступенів свободи V повинне бути додатним цілим числом. Значення ймовірності P повинне належати інтервалу [0 1]. Квантиль X є результатом розв'язання інтегрального рівняння рівного P з заданим числом ступенів свободи.

6.3. Розв'язування типової задачі в середовищі MATLAB

Задача. Дані статистичного щорічника ЮНЕСКО "Statistical Yearbook" про державні витрати на освіту й ВВП за 1984 р. наведені в табл. 6.2

Таблиця 6.2

Вихідні дані задачі

Країна	**Державні витрати на освіту (*EE*), млн дол.**	**Валовий внутрішній продукт (ВВП) (*GDP*), млд дол.**
1	2	3
Люксембург	0,34	5Д7
Уругвай	0,22	10,13
Сінгапур	0,32	11,34
Ірландія	1,23	18,88
Ізраїль	1,81	20,94
Угорщина	1,02	22,16
Нова Зеландія	1,27	23,83
Португалія	1,07	24,67
Гонконг	0,67	27,56
Чилі	1,25	27,57
Греція	0,75	40,15
Фінляндія	2,80	51,62
Норвегія	4,90	57,71
Югославія	3,50	63,03
Данія	4,45	66,32
Туреччина	1,60	66,97
Австрія	4,26	76,88
Швейцарія	5,31	101,65
Саудівська Аравія	6,40	115,97
Бельгія	7,15	119,49
Швеція	11,22	124,15
Австралія	8,66	140,98
Аргентина	5,56	153,85
Нідерланди	13,41	169,38
Мексика	5,46	186,33
Іспанія	4,79	211,78
Бразилія	8,92	249,72
Канада	18,90	261,41
Італія	15,95	395,52

Закінчення табл. 6.2

1	2	3
Великобританія	29,90	534,97
Франція	33,59	655,29
ФРН	38,62	815,00
Японія	61,61	1040,45
США	181,30	2586,40

Для початку роботи необхідно створити новий М-файл. Для цього з меню *File* треба вибрати опцію *New*, а потім M-File. У вікні, що з'явилося, редактора М-файлів ввести вихідні дані у вигляді масивів-стовпців. Використовуючи функцію *size*, визначити розмір вибірки (*n*) і кількість факторів, включених у модель (*m*).

Фрагмент М-файла

```
%% Гетероскедастичність
clc
clear all
%% Вихідні дані
Y = [0.34 0.22 0.32 1.23 1.81 1.02 1.27 1.07 0.67 1.25 0.75 2.80
4.90 3.50 4.45  1.60 4.26 5.31 6.40 7.15 11.22 8.66 5.56 13.41
5.46 4.79 8.92 18.90 15.95 29.90 33.59 38.62 61.61 181.30]';
X = [5.67 10.13 11.34 18.88 20.94 22.16 23.83 24.67 27.56 27.57
40.15 51.62 57.71 63.03 66.32 66.97 76.88 101.65 115.97 119.49
124.15 140.98 153.85 169.38 186.33 211.78 249.72 261.41 395.52
534.97 655.29 815.00 1040.45 2586.40]';
[n,m] = size(X);
```

За даними, наведеним у табл. 6.2, за допомогою функції *glmfit* оцінена регресійна залежність витрат на освіту (*EE*) від валового внутрішнього продукту (*GDP*). Графічне відображення вихідних даних та лінії регресії зображно на рис. 6.1.

Фрагмент М-файла

```
[P,dev,stats] = glmfit(X,Y);
y_p = P(1) + P(2).*X;
sprintf('Параметри моделі:')
P
fprintf('GDP_p = %f + %f *EE',P)
sprintf('Стандартні похибки:')
stats.se
sprintf('Коефіцієнт детермінації дорівнює')
```

```
Rxy = corrcoef(X,Y);
R2 = Rxy(1,2)^2
%R2 = sum((y_p - mean(Y)).^2)/sum((Y-mean(Y)).^2)
fprintf('Критерій Фішера')
F = (R2/(1-R2))*(n - m - 1)% F-статистика
Ft = finv(0.95,m,n-m-1)
if F > Ft
    fprintf('Модель значуща в цілому\n')
else
    fprintf('\nМодель не є значущой')
end
```

```
>> Параметри моделі:

P =

   -2.3199
    0.0669

GDP_p = -2.319896 + 0.066891 *EE

>> Стандартні похибки:

    0.9139
    0.0017

>> Коефіцієнт детермінації дорівнює

R2 =

    0.9794

>> Критерій Фішера

F =

  1.5245e+003
Ft =

    4.1491

Модель значуща в цілому
```

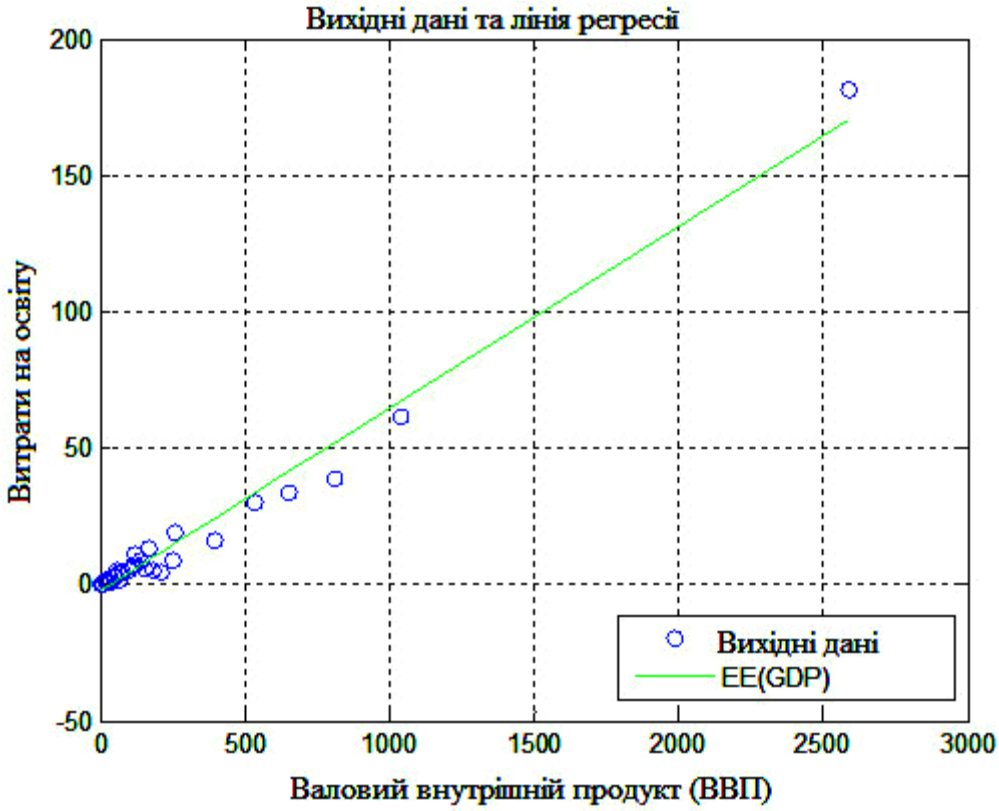


Рис. 6.1. **Вихідні дані та лінія регресії**

Це припускає, що з кожним збільшенням ВВП на 1 млрд дол. витрати на основі будуть збільшуватись на 67 млн дол.

Рівняння регресії значуще в цілому за критерієм Фішера – обчислене значення F = 152,45 значно перевищує табличне (Ft = 4,1491), знайдене за функцією `finv`. Коефіцієнт детермінації дорівнює 97,94 % , тобто витрати на освіту практично повністю визначається ВВП.

Необхідно обчислити значення відхилень і побудувати графічне подання отриманих відхилень (рис. 6.2).

Фрагмент М-файла

```
%% Графічний аналіз залишків:
e = Y - y_p;
plot(X, e.^2, 'ob')
grid on
xlabel('ВВП')
ylabel('Відхилення e = Y - Yp')
```

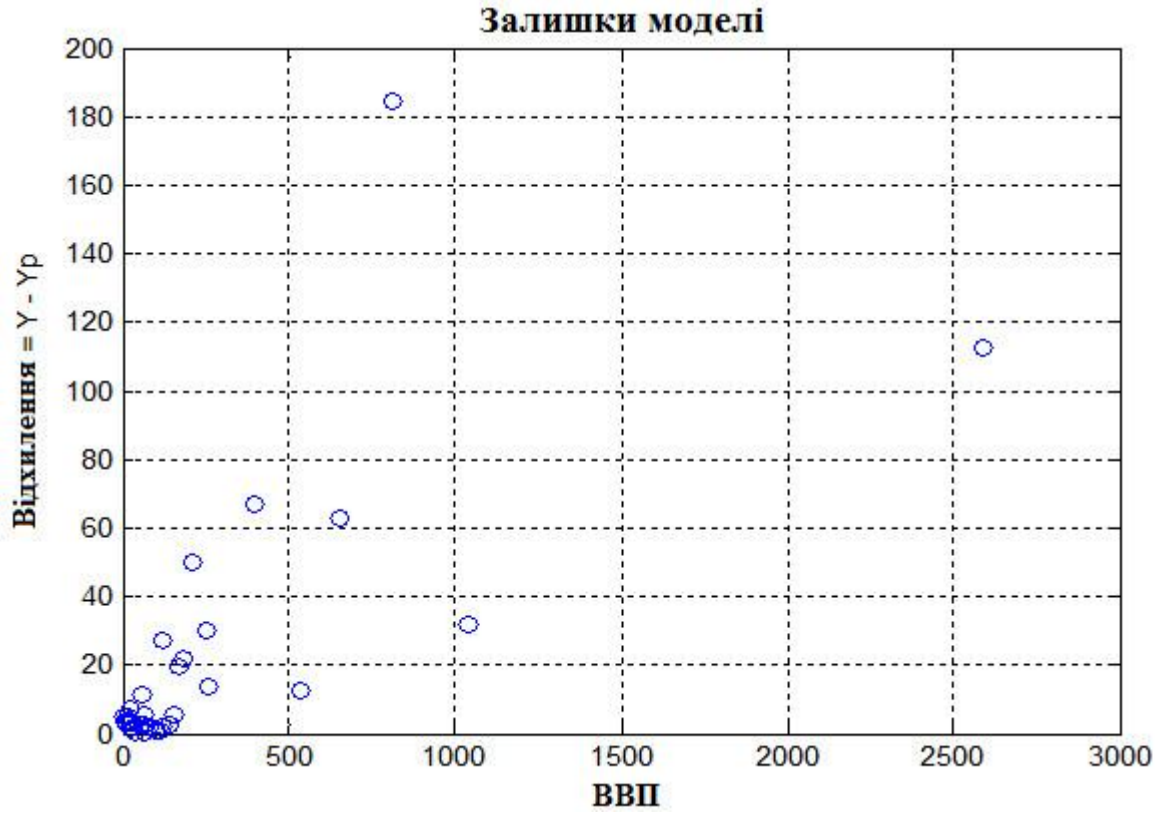


Рис. 6.2. **Графічний аналіз залишків**

Як видно із рис. 6.2, можна висунути гіпотезу про наявність гетероскедастичності. Однак це припущення необхідно підтвердити такими тестами, як тест Спірмена та Голдфельда – Квандта.

Програмна реалізація тесту Спірмена наведена в наступному фрагменті М-файла.

Фрагмент М-файла

```
%% Тест рангової кореляції Спірмена
% Ранжування:
[Xs I] = sort(X); Xs = [Xs ones(size(X),1)];
d = [zeros(size(X),1)]; X = [X ones(size(X),1)];
for i = 1:size(X)
    Xs(i,2) = i;
end
for i=1:size(X)
    X(I,2) = Xs(:,2);
end
[es I] = sort(abs(e)); es = [es ones(size(e),1)];
e = [e ones(size(Y),1)];
for i = 1:size(e)
```

```
    es(i,2) = i;
end
for i=1:size(e)
    e(I,2) = es(:,2);
end
for i=1:size(X)
    d(i) = (X(i,2) - e(i,2))^2;
end
X = X(:,1); e = e(:,1);
% Обчислення коефіцієнта Спірмена:
sprintf('коефіцієнт Спірмена:','\n')
r = 1 - 6*sum(d)/(n*(n^2 - 1))
%% Перевірка значущості коефіцієнта Спірмена
t_t = 1.96;
sprintf('критичне значення t:','\n')
t_r = r*sqrt(n-1)
if t_r < t_t
    sprintf('Гетероскедастичність відсутня')
else
    sprintf('Гетероскедастичність присутня')
end
```

```
>> Коефіцієнт Спірмена:

r = 0.5911

>> критичне значення t:

t_r = 3.3958

Гетероскедастичність присутня

>> k =

    13

>> S1 =

    6.0929

>> S2 =

  429.6117

Гетероскедастичність присутня
```

Програмна реалізація тесту Голдфельда – Квандта наведена в наступному фрагменті М-файла.

Фрагмент М-файла

```
%% Тест Голдфельда – Квандта :
[Xsort Is] = sort(X);
for i=1:size(Y)
    Ysort(i,1) = Y(Is(i),1);
end
Dat = [Xsort Ysort];
c = fix(4*n/15);
k = fix((n - c)/2);
if floor(k) > 0.4
    k = k+1;
end
k
% Вибірка 1:
Dat1 = Dat(1:k,:);
[b1,dev1,stats1] = glmfit(Dat1(:,1),Dat1(:,2));
S1 = sum(stats1.resid.^2);
% Вибірка 2:
Dat2 = Dat(n-k+1:n,:);
[b2,dev2,stats2] = glmfit(Dat2(:,1),Dat2(:,2));
S2 = sum(stats2.resid.^2);
% Перевірка гіпотези:
if S1 > S2
    Fp = S1/S2;
else
    Fp = S2/S1;
end
Ft = finv(0.95,k-m-1,k-m-1);
if Fp > Ft
     sprintf('Гетероскедастичність присутня')
else
    sprintf('Гетероскедастичність відсутня')
end
```

Обидва тести вказують на наявність гетероскедастичності в моделі. Для оцінювання параметрів регресійної моделі в умовах гетероскедастичності використовується функція *robustfit*. Для її використання потрібно здійснити ряд дій у *MATLAB*, наведених у наступному фрагменті М-файла.

Фрагмент М-файла

```
%% Розрахунок параметрів регресії в умовах гетероскедастичності:
[b,stats3] = robustfit(X,Y);
sprintf('Параметри моделі:')
b
fprintf('GDP_p = %f + %f *EE',b)
sprintf('Стандартні похибки:')
stats3.se
sprintf('Коефіцієнт детермінації дорівнює')
Rxy = corrcoef(X,Y);
R2 = Rxy(1,2)^2
%R2 = sum((y_p - mean(Y)).^2)/sum((Y-mean(Y)).^2)
fprintf('Критерій Фішера')
F = (R2/(1-R2))*(n - m - 1)% F-статистика
Ft = finv(0.95,m,n-m-1)
if F > Ft
    fprintf('Модель значуща в цілому\n')
else
    fprintf('\nМодель не є значущою')
end
```

```
>> Параметри моделі:
b =
   -0.2254
    0.0591

GDP_p = -0.225417 + 0.059064 *EE

>> Стандартні похибки:
    0.3917
    0.0007

>> Коефіцієнт детермінації дорівнює
R2 = 0.2185

>> Критерій Фішера
F =  8.9466
Ft = 4.1491

Модель значуща в цілому
```

Може викликати занепокоєння те, що рівень коефіцієнта детермінації R^2 став нижчим, ніж у першому рівнянні. Дійсно, в отриманому рівнянні він став набагато меншим, однак *F*-статистика вказує на значущість моделі в цілому.

Запитання для самоперевірки

1. Поясніть сутність гетероскедастичності.
2. Які відомі наслідки гетероскедастичності?
3. Наведіть методи визначення гетероскедастичності.
4. У чому сутність тесту рангової кореляції Спірмена?
5. У чому сутність тесту Парка?
6. У чому сутність тесту Голдфельда − Квандта?
7. Наведіть методи пом'якшення гетероскедастичності.
8. У чому сутність методу зважених найменших квадратів?
9. Чому за наявності гетероскедастичності ЗНК дозволяє отримати більш ефективні оцінки, ніж звичайний МНК?

Завдання для лабораторної роботи

Статистичні дані про виробництво електричної енергії на теплових електростанціях, а також її імпорт наведені в табл. 6.3 (p_1 – кількість букв у повному імені, p_2 – кількість букв у прізвищі).

Таблиця 6.3

Вихідні дані задачі

№ п/п	Виробництво на теплових електростанціях, мільйонів кіловат-годин, X	Імпорт, мільйонів кіловат-годин, Y
1	30924+ p_1	4936 + p_2
2	31775 – p_1	5479 + p_2
3	31793 – p_1	4344– p_2
5	30359+ p_1	4478– p_2
6	34844 – p_1	2971– p_2
7	32157 – p_1	5736 + p_2
8	30723+ p_1	7899 + p_2
9	31361+ p_1	6716 + p_2

За даними, що містяться в табл. 6.3, виконати наступні дії.

1. Побудувати модель регресії на основі статистичних даних.
2. Побудувати графічне подання відхилень. На підставі графічних даних висунути гіпотезу про наявність гетероскедастичності.

3. Провести аналітичну перевірку гетероскедастичності, використовуючи тест рангової кореляції Спірмена та тест Голдфельда – Квандта.

4. Оцінити параметри регресійної моделі в умовах гетероскедастичності.

5. Скориставшись зваженим методом найменших квадратів (WLS), зменшити гетероскедастичність.

6. Провести порівняльний аналіз отриманих моделей.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

Лабораторна робота 7
Автокореляція залишків. Критерій Дарбіна – Уотсона

Мета роботи: мати уяву про автокореляцію залишків та її наслідки. Вивчити методи виявлення автокореляції з використанням вбудованих функцій *MATLAB*.

Основні задачі лабораторної роботи:

1. Засобами *MATLAB* побудувати рівняння лінійної регресії.

2. Провести аналіз отриманої моделі.

3. Побудувати послідовно-часовий графік моделі та зробити припущення про наявність автокореляції.

4. Підтвердити наявність автокореляції в залишках за допомогою критерію Дарбіна – Уотсона та вбудованої функції `autocorr`. Порівняти отримані результати.

5. Обчислити коефіцієнт кореляції залишків першого порядку та перевірити справедливість співвідношення $DW \approx 2(1-\rho_1)$.

6. Усунути автокореляцію, використовуючи засоби *MATLAB*.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

7.1. Сутність автокореляції залишків

У динамічних рядах члени ряду пов'язані між собою: попередні члени впливають на наступні. Це явище називається ***автокореляцією***. Тому, перш ніж знаходити кількісну оцінку зв'язку між часовими рядами, необхідно перевірити існування автокореляції. Розрізняють автокореляцію змінних y і x_i і автокореляцію залишків моделі.

Важливою передумовою побудови якісної регресійної моделі за МНК є незалежність значень випадкових відхилень ε_i від значень відхилень у всіх інших спостереженнях. Відсутність залежності гарантує відсутність корельованості між будь-якими відхиленнями $\sigma(\varepsilon_i, \varepsilon_j) = 0$, якщо $i \neq j$ і, зокрема, між сусідніми відхиленнями $\sigma(\varepsilon_{i-1}, \varepsilon_i) = 0$.

Автокореляція (послідовна кореляція) визначається як кореляція між значеннями спостережуваних показників, впорядкованими в часі (часові ряди) або у просторі (перехресні дані). Автокореляція залишків (відхилень) зазвичай зустрічається в регресійному аналізі з використанням даних часових рядів. Поряд з від'ємною автокореляцією найчастіше зустрічається так звана додатна автокореляція. У більшості випадків додатна автокореляція обумовлюється спрямованою постійною дією деяких не врахованих у моделі факторів.

Серед основних причин, що викликають появу автокореляції, виділяють помилки специфікації, інерцію у зміні економічних показників, ефект павутини, згладжування даних. Неврахування в моделі важливої змінної або неправильний вибір форми залежності призводить до системних відхилень точок спостережень від ліній регресії, що може зумовити автокореляцію.

Наслідки автокореляції аналогічні наслідкам гетероскедастичності. Зазвичай виокремлюють наступні наслідки.

1. Оцінки параметрів, залишаючись лінійними і незміщеними, не є ефективними. Отже, вони перестають мати властивості найкращих лінійних незміщених оцінок (BLUE-оцінок).

2. Дисперсії оцінок будуть зміщеними. Часто дисперсії, які розраховуються за стандартними формулами, є заниженими. Це призводить до збільшення *t*-статистик. Що, у свою чергу, тягне за собою визнання значущими пояснювальних змінних, які такими не є. Висновки за *t*- і *F*-статистикою можуть бути неправильними, що погіршує статистичну якість моделі.

3. Оцінка дисперсії регресії $S^2 = \sum_{t=1}^{T} \frac{\varepsilon_t^2}{T - m - 1}$ є незміщеною оцінкою справжнього значення σ^2, у багатьох випадках занижуючи її. Тут $T = n, t = i$.

4. Підсумок наслідків попередніх розрахунків доводить, що висновки за *t*- і *F*-статистикою, які визначають значущість коефіцієнтів регресії і коефіцієнта детермінації, можуть бути неправильними, що погіршує прогнозні якості моделі.

Існує кілька методів виявлення автокореляції.

Графічний метод. На графік реальних коливань залежної змінної накладається графік коливань змінної за рівнянням регресії. Порівнюючи два

графіка, висувають гіпотезу про наявність автокореляції залишків. Якщо ці графіки перетинаються рідко, то можна передбачити існування додатної автокореляції залишків.

Метод рядів. У даному методі послідовно визначають знаки відхилень $\varepsilon_t, t=\overline{1,T}$. Наприклад,

(- - - - -)(+ + + + + + +)(- - -)(+ + + +)(-),

тоді 5 "-", 7 "+", 3 "-", 4 "+", 1 "-" з 20-ти спостережень.

Ряд є неперервна послідовність однакових знаків, при цьому довжина ряда визначається кількістю знаків у ряді. Якщо рядів замало порівняно з кількістю спостережень n, то ймовірна додатна автокореляція. Якщо ж рядів дуже багато, то ймовірна від'ємна автокореляція.

Нехай n – обсяг вибірки, n_1 – загальна кількість знаків "+" в n спостереженнях, n_2 – загальна кількість знаків "-" в n спостереженнях, k – кількість рядів. Для $n_1 > 10$ і $n_2 > 10$ та за відсутності автокореляції випадкової величини k має місце асимптотично нормальний розподіл з:

$$M(k) = \frac{2n_1 n_2}{n_1 + n_2} + 1; \quad D(k) = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 + 1)} + 1. \tag{7.1}$$

Якщо $M(k) - u_{\frac{\alpha}{2}} D(k) < k < M(k) + u_{\frac{\alpha}{2}} D(k)$, то гіпотеза про відсутність автокореляції не відхиляється.

Для невеликої кількості спостережень $n_1 < 20$ і $n_2 < 20$ вчені Свед і Ейзенхарт розробили таблиці критичних значень кількості рядів для n спостережень. З таблиць визначають нижнє k_1 і верхнє k_2 значення за рівнем значущості $\alpha = 0{,}05$. Якщо $k_1 < k < k_2$, то говорять про відсутність автокореляції; якщо $k < k_1$, то говорять про додатну автокореляцію; якщо $k > k_2$, то говорять про від'ємну автокореляцію залишків.

Критерій Дарбіна – Уотсона. Найбільш відомим критерієм виявлення автокореляції першого порядку є критерій Дарбіна – Уотсона. На основі обчисленої статистики DW Дарбіна – Уотсона здійснюють висновок про автокореляцію:

$$DW = \frac{\sum_{t=2}^{T} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{T} \varepsilon_t^2}. \tag{7.2}$$

Таким чином, статистика Дарбіна − Уотсона тісно пов'язана з вибірковим коефіцієнтом кореляції $r_{\varepsilon_t\varepsilon_{t-1}}$: $DW \approx 2(1 - r_{\varepsilon_t\varepsilon_{t-1}})$. Якщо $0 \le DW \le 4$, то її значення можуть вказати на наявність або відсутність автокореляції. Якщо $r_{\varepsilon_t\varepsilon_{t-1}} = 0$ (автокореляція відсутня), то $DW \approx 2$. Якщо $r_{\varepsilon_t\varepsilon_{t-1}} = 1$ (додатна автокореляція), то $DW \approx 0$. Якщо $r_{\varepsilon_t\varepsilon_{t-1}} = -1$ (від'ємна автокореляція), то $DW \approx 4$. У критерії Дарбіна − Уотсона з заданим рівнем значущості α, кількості спостережень n і кількості пояснювальних змінних m визначаються два значення: d_1 – нижня межа і d_u – верхня межа. Загальна схема критерію Дарбіна − Уотсона наступна.

1. Для побудованого емпіричного рівняння регресії $\hat{y} = a + b_1x_1 + b_2x_2 + ... + b_mx_m$ визначають значення відхилень $\varepsilon_t = y_t - \hat{y}_t$.

2. За формулою

$$DW = \frac{\sum_{t=2}^{T}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{T}\varepsilon_t^2} \qquad (7.3)$$

розраховують статистику DW.

3. За таблицею критичних точок Дарбіна − Уотсона визначають два числа d_l і d_u, потім роблять висновки за правилом:

$0 \le DW \le d_l$ – існує додатна кореляція;

$d_l \le DW \le d_u$ – висновок про наявність автокореляції не визначений;

$d_u \le DW \le 4 - d_u$ – автокореляція відсутня;

$4 - d_u \le DW \le 4 - d_l$ – висновок про наявність автокореляції не визначений;

$4 - d_l \le DW \le 4$ – існує від'ємна автокореляція.

У процесі використання критерія Дарбіна − Уотсона слід враховувати обмеження: 1) критерій Дарбіна − Уотсона застосовується лише для моделей з вільним членом; 2) передбачається, що випадкові відхилення ε_t визначаються за ітераційною схемою $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$, яку називають авторегресійною схемою першого порядку $AR(1)$, де v_t – випадковий член; 3) статистичні дані повинні мати однакову періодичність; 4) критерій Дарбіна − Уотсона не слід застосовувати для регресійних моделей, що містять в складі пояснюючих змінних лагову змінну y_{t-1}, тобто для випадку:

$$y = a + b_1x_{t1} + b_2x_2 + ... + b_mx_{t\,m} + \gamma y_{t-1} + \varepsilon_t. \qquad (7.4)$$

Для авторегресійних моделей розроблені спеціальні тести виявлення автокореляції, зокрема *h*-статистика Дарбіна, яку обчислюють за формулою:

$$h = \hat{\rho}\sqrt{\frac{n}{1 - nD(\gamma)}}, \tag{7.5}$$

де $\hat{\rho}$ – оцінка ρ авторегресії першого порядку;

$D(\gamma)$ – вибіркова дисперсія коефіцієнта з лаговою змінною y_{t-1};

n – кількість спостережень.

Для великого n справедлива нульова гіпотеза H_0: $\rho = 0$ статистика h має стандартизований нормальний розподіл ($h \sim N(0,1)$).

Для заданого рівня значущості α визначають критичну точку $u_{\frac{\alpha}{2}}$ з умови $\Phi\left(u_{\frac{\alpha}{2}}\right) = \frac{1-\alpha}{2}$ і порівнюють h з $u_{\frac{\alpha}{2}}$. Якщо $|h| > u_{\frac{\alpha}{2}}$, то нульову гіпотезу про відсутність автокореляції відхиляють, у протилежному випадку вона не відхиляється.

Методи усунення автокореляції. Основною причиною наявності випадкового члена в моделі є недосконалі знання про причини і взаємозв'язки, що визначають те чи інше значення залежної змінної. Тому властивості випадкових відхилень, в тому числі й автокореляція, в першу чергу залежать від вибору формули залежності та складу пояснювальних змінних. Оскільки автокореляція найчастіше викликана неправильною специфікацією моделі, то необхідно насамперед скорегувати саму модель. Також можна спробувати змінити форму залежності (наприклад, лінійну на лог-лінійну або гіперболічну). Можна скористатися авторегресійним перетворенням, а саме – авторегресійною схемою першого порядку *AR*(1).

Якщо залишки вихідного рівняння регресії містять автокореляцію, то для оцінювання параметрів рівняння використовують *узагальнений метод найменших квадратів*. Для його реалізації необхідно виконувати наступні умови:

1. Перетворити вихідні змінні *y* і *x* до вигляду:

$$\begin{aligned} y'_i &= y_i - \rho_1 y_{i-1}, \\ x'_i &= x_i - \rho_1 x_{i-1}. \end{aligned} \tag{7.6}$$

2. Застосувавши звичайний МНК до рівняння $\tilde{y}'_p = b'_0 + b'_1 x' + \varepsilon'$, визначити оцінки параметрів b_0' й b_1'.

3. Розрахувати параметр b_0 вихідного рівняння як:

$$b_0 = \frac{b'_0}{1-\rho_1}. \tag{7.7}$$

4. Виписати вихідне рівняння лінійної регресії.

7.2. Теоретичні відомості про функції *MATLAB*, які використовуються в даній лабораторній роботі

AUTOCORR **Функція автокореляції (АКФ)**

Синтаксис

```
autocorr(Series,nLags,M,nSTDs)
[ACF,Lags,Bounds] = autocorr(Series,nLags,M,nSTDs)
```

Опис

`autocorr(Series,nLags,M,nSTDs)` обчислює та зображує АКФ одномірного, стохастичного часового ряду з довірчими інтервалами. Для побудови послідовності АКФ без довірчих інтервалів, необхідно встановити `nSTDs = 0`.

`[ACF,Lags,Bounds] = autocorr(Series,nLags,M,nSTDs)` обчислює та повертає послідовність АКФ.

У табл. 7.1 наведено опис усіх вхідних параметрів функції `autocorr.`

Таблиця 7.1

Вхідні та вихідні аргументи функції *autocorr*

Назва параметру	Значення параметру
1	2
`Series`	Вектор-стовпець спостережень одномірних часових рядів, для яких *autocorr* обчислює та графічно зображує ділянки функції автокореляції (АКФ). Останній рядок `series` містить останні спостереження часового ряду

Закінчення табл. 7.1

1	2
nLags	Додатній скаляр (ціле число), що вказує число лагів АКФ для обчислення. Якщо *nLags* = [] або не заданий, то за замовчуванням обчислюється АКФ наступних лагів: 0, 1, 2, ..., T, де *T = min*([20,*length*(*Series*)-1])
M	Від'ємне ціле число, що вказує число лагів, за яке теоретичне АКФ стає ефективним. *autocorr* реалізує основний MA (M) процес і використовує апроксимацію Бартлета для обчислення стандартної похибки для запізнювання, більшого ніж *M*. Якщо *M* = [] або не задано, то за замовчуванням воно дорівнює 0, і *autocorr* допускає, що має місце гауссовий білий шум. Якщо часовий ряд містить гауссовий білий шум довжини *N*, стандартна похибка обчислюється як $\frac{1}{\sqrt{N}}$. *M* повинне бути менше, ніж *nLags*
nSTDs	Додатній скаляр із вказаною кількістю стандартних відхилень від обчисленого АКФ (похибка обчислення). *autocorr* надає теоретичному ряду АКФ нульове значення лага *M*. Якщо *M* = 0, часовий ряд, що має гаусовий білий шум довжини *N*, у параметрі *nSTDs* містить довірчі межі в межах $\pm\frac{nSTDs}{\sqrt{N}}$. Якщо *nSTDs* = [] або не заданий, за замовчуванням йому надається значення 2 (тобто, приблизно 95-відсотковий довірчий інтервал).
ACF	Автокореляціна функція часового ряду. АКФ є вектором довжиною *nLags* + 1 з відповідними лагами 0, 1, 2, ..., *nLags*. Перший елемент АКФ є одиницею, тобто, АКФ (1) = 1 = *lag* 0
Lags	Вектор лагів, що відповідає АКФ (0,1,2, ..., *nLags*). АКФ симетрична щодо нульового лага, *autocorr* ігнорує від'ємні лаги
Bounds	Вектор, що складається із двох елементів, містить наближення верхніх й нижніх меж довірчих інтервалів. *autocorr* обчислює межі тільки для лагів, більших M

7.3. Розв'язування типової задачі в середовищі *MATLAB*

Задача. Нехай є дані про середньодушовий дохід і середньодушові витрати на кінцеве споживання в США за період 1960 – 1991 рр. (табл. 7.2).

Таблиця 7.2

Вихідні дані задачі

Рік	**Середньодушові витрати на кінцеве споживання (дол. США),** *y*	**Середньодушовий доход (дол. США)**, *x*
1960	6 698	7 264
1961	6 740	7 382
1962	6 931	7 583
1963	7 089	7 718
1964	7 384	8 140
1965	7 703	8 508
1966	8 005	8 822
1967	8 163	9 114
1968	8 506	9 399
1969	8 737	9 606
1970	8 842	9 875
1971	9 022	10 111
1972	9 425	10 414
1973	9 752	11 013
1974	9 602	10 832
1975	9 711	10 906
1976	10 121	11 192
1977	10 425	11 406
1978	10 744	11 851
1979	10 867	12 039
1980	10 746	12 005
1981	10 770	12 156
1982	10 782	12 146
1983	11 179	12 349
1984	11 617	13 029
1985	12 015	13 258
1986	12 336	13 552
1987	12 568	13 545
1988	12 903	13 890
1989	13 027	14 030
1990	13 051	14 154
1991	12 889	13 987

Для початку роботи необхідно створити новий M-файл. Для цього з меню File вибрати опцію New, а потім M-File. У вікні, що з'явилося, редактора M-файлів ввести вихідні дані. Це можна зробити двома способами: отримати дані з файлу або ввести дані вручну.

Фрагмент М-файла

```
clc
clear all
x = [7264 7382 7583 7718 8140 8508 8822 9114 9399 9606 9875 10111
10414 11013 10832 10906 11192 11406 11851 12039 12005 12156 12146
12349 13029 13258 13552 13545 13890 14030 14154 13987]';
y = [6698 6740 6931 7089 7384 7703 8005 8163 8506 8737 8842 9022
9425 9752 9602 9711 10121 10425 10744 10867 10746 10770 10782
11179 11617 12015 12336 12568 12903 13027 13051 12889]';
[n, m] = size(x);
```

В останньому рядку наведеного програмного коду автоматично визначаються обсяг вибірки (n) і кількість факторів (m).

Засобами *MATLAB* побудувати рівняння лінійної регресії $\tilde{y}_p = b_0 + b_1 x$; провести аналіз отриманої моделі та надати економічну інтерпретацію отриманих даних. Для обчислення моделі використаємо вбудовані функції *glmfit* та *regress*.

Фрагмент М-файла

```
%% Побудуємо рівняння лінійної регресії:
[b,dev,stats] = glmfit(x,y);
b0 = b(1);
b1 = b(2);
yp = b0 + b1*x;% Формування масиву розрахункових даних
fprintf('\nРівняння регресії\n')
fprintf('\n yp = %f + %f *x',b)
fprintf('\nСтатистики Стьюдента:')
for i = 1:(m+1)
    tb(i) = b(i)/stats.se(i);
end
tb
%% Коефіцієнт детермінації і критерій Фішера
X = [ones(n,1) x];
[b,bint,r,rint,stats] = regress(y,X,0.05);
R2 = stats(1)% коефіцієнт детермінації
F = stats(2)%F-статистика
Ft = finv(0.95,m,n-m-1);
if F > Ft
    fprintf('\nМодель значуща в цілому\n')
else
    fprintf('\nМодель не є значущою\n')
end
%% Значущість параметрів моделі з використанням критерію Стьюдента:
tt = tinv(0.95,n-m-1);
```

```
for i = 1:m+1
    if abs(tb(i)) > tt
        fprintf('\nПараметр %f є значущим\n',b(i))
    else
        fprintf('\nПараметр %f не є значущим\n',b(i))
    end
end
```

```
>> Рівняння регресії

 yp = -174.304974 + 0.922146 *x

>> Статистики Стьюдента:

tb =   -1.2134   71.7841
>> R2 =

    0.9942

>> F =

  5.1530e+003

Модель значуща в цілому

Параметр -174.304974 не є значущим

Параметр 0.922146 є значущим
```

Таким чином, було отримано рівняння лінійної регресії (у дужках наведені значення *t*-статистик Стьюдента):

$$\tilde{y}_p = -174.31 + 0.92x. \quad (7.8)$$
$$(-1.21) \quad (71.78)$$

Коефіцієнт b_1 показує, що зі зміною середньодушового розпоряджуваного доходу на 1 дол. США середньодушові витрати на кінцеве споживання збільшаться в середньому на 0,92 дол. США (на 92 цента).

Коефіцієнт детермінації R^2 показує, що 99,42 % загальної зміненюваності середньодушових витрат кінцевого споживання обумовлене середньодушовим доходом (дол. США).

Дисперсійне відношення Фішера F = 5153 показує, що змінюваність обчислених значень майже в 5153 разів перевищує зміненюваність залишків моделі. Порівнянням обчисленого значення F з табличним було виявлено, що регресійна модель значуща в цілому.

Отримана регресійна модель доповнена рядком відношень $t_b = b/S_b$, які показують, у скільки разів обчислені значення параметрів перевершують свої стандартні похибки. Порівнянням цих статистик Стьюдента з табличними значеннями встановлено, що параметр b_0 не є значущим, а b_1 є значущим.

Використовуючи вбудовану функцію `plot`, слід побудувати послідовно-часовий графік моделі (рис. 7.1), на підставі якого можна буде зробити припущення про наявність автокореляції.

Фрагмент М-файла

```
%% Послідовно-часовий графік:
i = zeros(n,1);
for j = 1:n
    i(j) = i(j)+j;
end
plot(i,r,'bo-')
grid on
title('Послідовно-часовий графік')
xlabel('t')
ylabel('et')
```

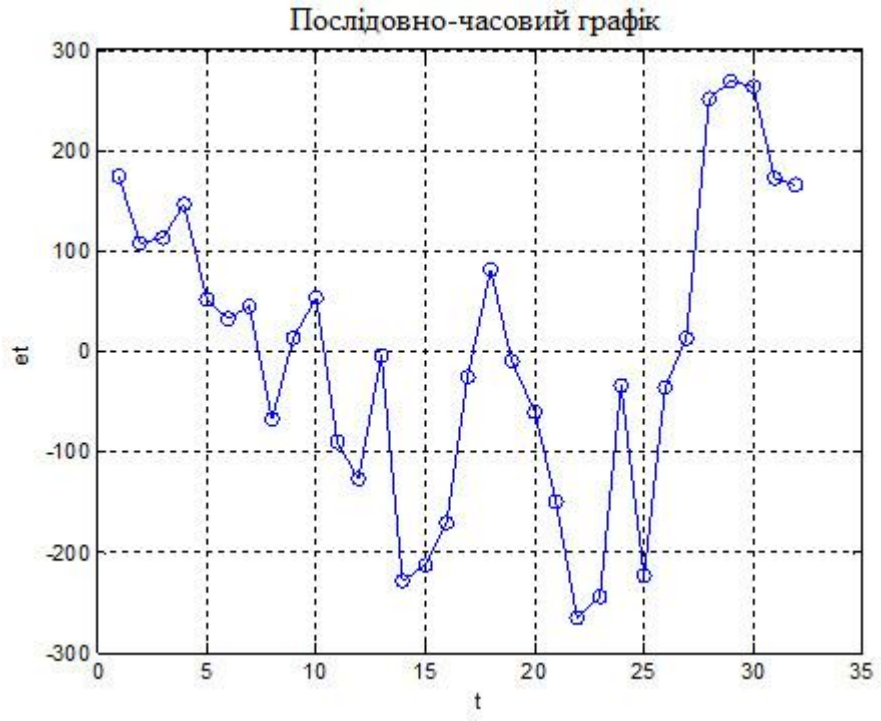


Рис. 7.1. **Послідовно-часовий графік моделі**

Із зовнішнього вигляду послідовно-часового графіка можна зробити припущення про наявність автокореляції в залишках даної моделі.

Необхідно перевірити гіпотезу про наявність автокореляції в залишках для моделі залежності середньодушових витрат на кінцеве споживання від середньодушового передбачуваного доходу. Для цього треба обчислити за формулою (7.3) критерій Дарбіна – Уотсона. Необхідно звернути увагу на те, що значення змінних *du* та *dl* необхідно задавати вручну, використовуючи таблицю статистик Дарбіна – Уотсона:

Фрагмент М-файла

```
%% Критерій Дарбіна-Уотсона:
s = 0;
for i = 2:n
    s = s + (r(i)-r(i-1)).^2;
end
chisl_DW = s;
znam_DW = sum(r.^2);
fprintf('\n Критерій Дарбіна – Уотсона дорівнює:\n')
DW = chisl_DW/znam_DW
dl = 1.35; %n = 32, m = 1
du = 1.49; %n = 32, m = 1
if DW >= 0 & DW<= dl
    fprintf('Існує додатна кореляція \ n')
end
if DW >= dl& DW <= du
    fprintf('Висновок про наявність автокореляції не
визначений\n')
end
if DW >= du & DW <= 4-du
    fprintf('Автокореляція відсутня\n')
end
if DW >= 4-du & DW <= 4-dl
    fprintf('Висновок про наявність автокореляції не
визначений\n')
end
if DW >= 4-dl & DW <= 4
    fprintf('Існує від'ємна автокореляція\n')
end
```

```
>> Критерій Дарбіна-Уотсона дорівнює:

DW =
    0.5193

Існує додатна кореляція
```

Таким чином, фактичне значення критерію Дарбіна – Уотсона для цієї моделі становить: 0,5 193. За таблицями значень критерію Дарбіна – Уотсона визначити для кількості спостережень n = 32, кількість незалежних змінних моделі m = 1 і рівня значущості 0,05 критичні значення dl = 1,35 і du = 1,49. Фактичне значення DW = 0,5 193 попадає в проміжок від 0 до dl, що свідчить про наявність додатньої автокореляції. Отже, гіпотезу H_0 про відсутність автокореляції в залишках необхідно відхилити.

Отримані результати повністю співпадають з роботою вбудованої функції `autocorr` (рис. 7.2).

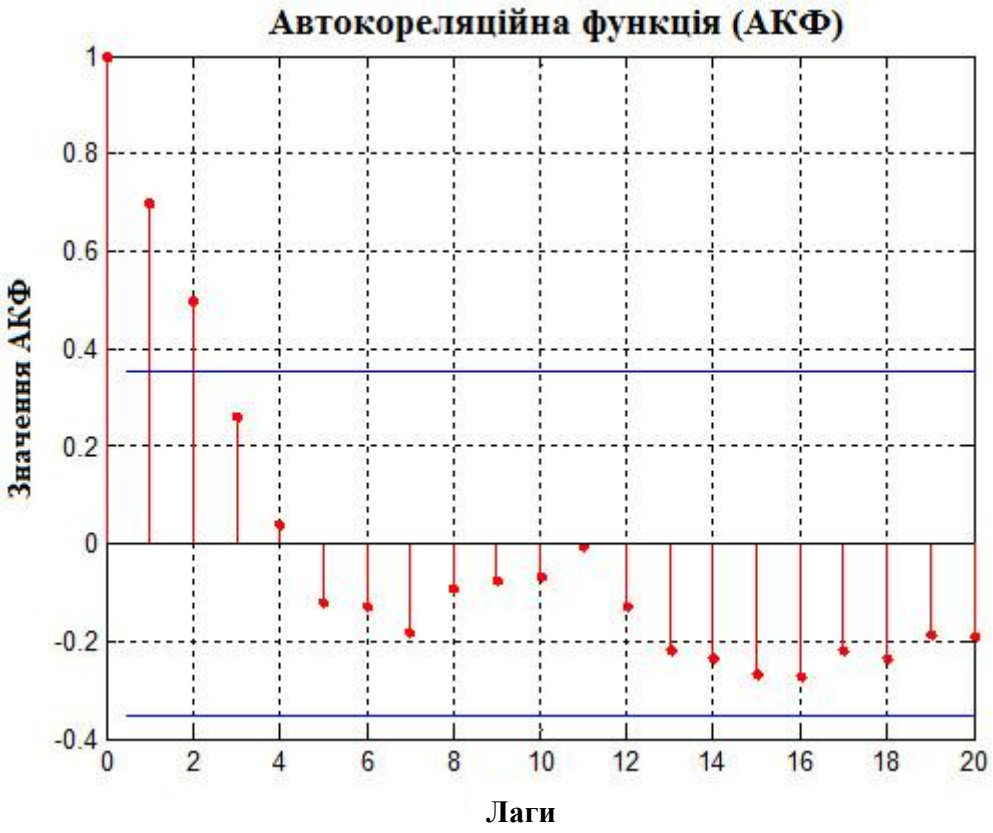


Рис. 7.2. **Графічне відображення автокореляційної функції**

Оцінку коефіцієнта автокореляції залишків першого порядку ρ_1 знаходять, використовуючи формулу (7.5). Необхідно порівняти його зі значенням критерію Дарбіна – Уотсона: має виконуватися співвідношення $DW \approx 2(1 - r_{\varepsilon_t \varepsilon_{t-1}})$.

Фрагмент М-файла

```
%% Коефіцієнт автокореляції першого порядку
ro1 = 0;
for i = 2:n
```

```
    ro1 = ro1 + r(i)*r(i-1);
end
ro1 = ro1/sum(r.^2)
fprintf('\Перевірка формули DW = 2*(1-ro1):\n')

DW
2*(1-ro1)
```

```
>> ro1 =

    0.7003
>> Перевірка формули DW = 2*(1-ro1):

DW =

    0.5193
ans =

    0.5994
```

Статистика Дарбіна – Уотсона тісно пов'язана з вибірковим коефіцієнтом кореляції ρ_1: $DW \approx 2(1-\rho_1)$.

Знайдені оцінки параметрів рівняння регресії b_0 = -174,305 і b_1 = 0,922 не є ефективними через порушення передумов МНК.

Для отримання нових оцінок параметрів, для яких не порушується властивість ефективності, використовують узагальнений метод найменших квадратів для обчислення параметрів рівняння регресії за наявності автокореляції в залишках.

Для програмної реалізації узагальненого методу найменших квадратів необхідно створити два нових масиви даних (*yn* і *xn*). Потім знайти оцінку коефіцієнта автокореляції залишків першого порядку, наприклад, скориставшись наближеним співвідношенням між критерієм Дарбіна – Уотсона та коефіцієнтом автокореляції залишків першого порядку. А після здійснити перетворення вихідних даних відповідно до формули (7.6):

Фрагмент М-файла

```
%% Узагальнений МНК:
ro1 = 1 - DW/2
for i=2:n
    yn(i,1) = y(i) - ro1*y(i-1);
```

```
    xn(i,1) = x(i) - ro1*x(i-1);
end
```

```
>> ro1 =

    0.7403
```

Майже такий же результат можна отримати, якщо обчислити коефіцієнт автокореляції рівнів першого порядку за рядом залишків ($\rho_1 = 0{,}729$). Нижче наведені вихідні дані та нові змінні *x* ' й *y* ':

```
>> [x y xn yn] =

  1.0e+004 *

        0.7264    0.6698         0         0
        0.7382    0.6740    0.2004    0.1781
        0.7583    0.6931    0.2118    0.1941
        0.7718    0.7089    0.2104    0.1958
        0.8140    0.7384    0.2426    0.2136
        0.8508    0.7703    0.2482    0.2236
        0.8822    0.8005    0.2523    0.2302
        0.9114    0.8163    0.2583    0.2237
        0.9399    0.8506    0.2652    0.2463
        0.9606    0.8737    0.2648    0.2440
        0.9875    0.8842    0.2763    0.2374
        1.0111    0.9022    0.2800    0.2476
        1.0414    0.9425    0.2929    0.2746
        1.1013    0.9752    0.3303    0.2774
        1.0832    0.9602    0.2679    0.2382
        1.0906    0.9711    0.2887    0.2602
        1.1192    1.0121    0.3118    0.2932
        1.1406    1.0425    0.3120    0.2932
        1.1851    1.0744    0.3407    0.3026
        1.2039    1.0867    0.3265    0.2913
        1.2005    1.0746    0.3092    0.2701
        1.2156    1.0770    0.3268    0.2814
        1.2146    1.0782    0.3147    0.2809
        1.2349    1.1179    0.3357    0.3197
        1.3029    1.1617    0.3887    0.3341
        1.3258    1.2015    0.3612    0.3415
        1.3552    1.2336    0.3737    0.3441
        1.3545    1.2568    0.3512    0.3435
```

```
    1.3890    1.2903    0.3862    0.3599
    1.4030    1.3027    0.3747    0.3475
    1.4154    1.3051    0.3767    0.3407
    1.3987    1.2889    0.3508    0.3227
```

Визначити параметри рівняння регресії *y*' на *x*' можна звичайним МНК.

Фрагмент М-файла

```
Xn = [ones(n,1) xn];
[b,bint,r,rint,stats] = regress(yn,Xn,0.05);
yp_n = b(1) + b(2)*xn;% формування масиву розрахункових даних
fprintf('\nРівняння регресії\n')
fprintf('\n yp_n = %f + %f *x_n\n',b)
```

```
>> Рівняння регресії

 yp_n = -83.797283 + 0.934267 *x_n
```

Для обчислення параметра b_0 вихідного рівняння необхідно скористатися формулою (7.6), тоді рівняння регресії залежності середньодушових витрат на кінцеве споживання від середньодушового розпоряджуваного доходу має вигляд:

Фрагмент М-файла

```
b(1) = b(1)/(1-ro1);
yp_n = b(1) + b(2)*xn;
fprintf('\nРівняння регресії\n')
fprintf('\n yp_n = %f + %f *x_n\n',b)
e = yn - yp_n;
R2 = 1-sum(e.^2)/sum((yn - mean(yn)).^2)
F = (R2/(1-R2))*(n - m - 1);% F-статистика
Ft = finv(0.95,m,n-m-1);
if F > Ft
    fprintf('\nМодель значуща в цілому\n')
else
    fprintf('\nМодель не є значущою\n')
end
```

```
>> Рівняння регресії

 yp_n = -322.703907 + 0.934267 *x_n
```

```
>> R2 =

    0.7353
```

```
Модель значуща в цілому
```

Отже, гранична схильність до споживання в США за період з 1960 по 1991 р. дорівнює 0,934. Це означає, що зі збільшенням середньодушового розпоряджуваного доходу на 1 дол. США середньодушові витрати на кінцеве споживання зростали в середньому на 93,4 цента.

Наостанок, необхідно провести аналіз параметрів моделі з різними значеннями ρ_1 (табл. 7.3).

З табл. 7.3 видно, що з урахуванням коефіцієнта автокореляції рівнів ряду першого порядку за залишками модель на певній ітерації покращується.

Таблиця 7.3

Значення параметрів моделі залежно від коефіцієнта автокореляції рівнів ряду

ρ_1	0	0,2	0,4	0,6	0,8
b_0	-236, 3 718	-256, 3 618	-285, 0 088	-322, 8 446	-260, 9 571
b_1	0, 9 272	0, 9 289	0, 9 314	0, 9 346	0, 9 288
R^2	0, 9 940	0, 9 918	0, 9 804	0, 9 178	0, 6 581
ρ_1^*	0, 6 983	0, 6 084	0, 6 509	0, 7 665	0, 7 608

Таким чином, дані обчислень показали ефективність узагальненого методу найменших квадратів.

Запитання для самоперевірки

1. Що таке автокореляція?
2. Які види автокореляції розрізняють?
3. У яких даних найчастіше зустрічається автокореляція?
4. Які основні причини автокореляції?
5. Які передумови МНК порушуються в автокореляції?

6. Які наслідки автокореляції?

7. Назвіть основні методи виявлення автокореляції.

8. Опишіть схему використання статистики Дарбіна − Уотсона.

9. Назвіть методи усунення автокореляції?

10. Поясніть особливості алгоритму узагальненого методу найменших квадратів.

Завдання для лабораторної роботи

Задача. Статистичні дані, які характеризують залежність обсягів продажів компанії за місяць *y* від витрат на рекламу *x* (у. е), наведені в табл. 7.4 (p_1/10 – кількість букв у повному імені, p_2/10 – кількість букв у прізвищі).

Таблиця 7.4

Вихідні дані задачі

Місяць	Витрати на рекламу, *x*	Обсяги продажів, *y*
1	$1{,}52 + p_1$	$17{,}955 + p_2$
2	$1{,}23 + p_1$	$16{,}842 + p_2$
3	$1{,}78 - p_1$	$18{,}448 - p_2$
4	$1{,}45 + p_1$	$18{,}286 - p_2$
5	$1{,}48 + p_1$	$17{,}388 - p_2$
6	$1{,}12 + p_1$	$16{,}345 + p_2$
7	$1{,}34 + p_1$	$15{,}029 + p_2$
8	$1{,}38 + p_1$	$15{,}626 + p_2$
9	$1{,}45 - p_1$	$16{,}241 + p_2$
10	$1{,}14 + p_1$	$16{,}100 + p_2$
11	$1{,}58 - p_1$	$17{,}446 + p_2$
12	$1{,}45 + p_1$	$17{,}351 - p_2$
13	$1{,}24 + p_1$	$16{,}872 + p_2$
14	$1{,}35 + p_1$	$16{,}087 + p_2$
15	$1{,}68 - p_1$	$17{,}496 - p_2$
16	$1{,}65 - p_1$	$18{,}844 - p_2$
17	$1{,}56 + p_1$	$18{,}904 - p_2$
18	$1{,}50 + p_1$	$19{,}214 - p_2$

За даними табл. 7.4 виконати наступні дії.

1. Засобами *MATLAB* побудувати рівняння лінійної регресії.

2. Провести аналіз отриманої моделі.

3. Побудувати послідовно-часовий графік моделі й зробити припущення про наявність автокореляції.

4. Підтвердити наявність автокореляції в залишках за допомогою критерію Дарбіна – Уотсона та вбудованої функції `autocorr`. Порівняти отримані результати.

5. Обчислити коефіцієнт кореляції залишків першого порядку та перевірити справедливість співвідношення $DW \approx 2(1-\rho_1)$.

6. Усунути автокореляцію, використовуючи засобами *MATLAB*.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

Лабораторна робота 8
Системи лінійних одночасних рівнянь

Мета роботи: мати уявлення про системи одночасних рівнянь, знати й уміти застосовувати непрямий та двокроковий методи найменших квадратів на основі проведеної ідентифікації моделі.

Основні задачі лабораторної роботи

1. За вихідним даними скласти структурну форму моделі.

2. Перевірити модель на ідентифікацію.

3. Непрямим методом найменших квадратів оцінити параметри структурної моделі.

3.1. Скласти приведену форму моделі.

3.2. Для кожного рівняння приведеної форми стандартним МНК визначити його коефіцієнти.

3.3. Перейти від приведеної до структурної форми моделі.

3.4. Оцінити кожне рівняння моделі, використовуючи *F*-критерій та коефіцієнт детермінації. Дати економічну інтерпретацію отриманих результатів.

4. Побудувати надідентифіковану модель, виходячи з попередньої моделі.

5. Двокроковим методом найменших квадратів оцінити параметри моделі.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

8.1. Основні поняття системи одночасних рівнянь

Використання одного окремого рівняння регресії для моделювання залежності результівної ознаки від факторів є грубим, реально в економіці існують механізми взаємозв'язків між факторами і результатами. Більш того, в одному випадку ознака є фактором, в іншому є результатом. В одних рівняннях

певна змінна розглядається як пояснювальна (незалежна), в інші рівняння системи вона входить як залежна змінна. Тобто на практиці для моделювання економічних явищ, процесів слід будувати таку модель:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + ... + b_{1n}y_n + a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + ... + b_{2n}y_n + a_{21}x_1 + a_{22}x_2 + ... + a_{2m}x_m + \varepsilon_1, \\ ..... \\ y_n = b_{n1}y_1 + b_{n2}y_2 + ... + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + ... + a_{nm}x_m + \varepsilon_n. \end{cases} \quad (8.1)$$

Для опису структури зв'язків між змінними використовують систему так званих одночасних рівнянь, які ще називають структурними рівняннями (8.1). В економетриці цю систему рівнянь називають структурною формою моделі. Кожне рівняння системи одночасних рівнянь не може розглядатись самостійно. Для знаходження параметрів системи одночасних рівнянь класичний МНК не застосовується.

Одним з найпростіших прикладів систем одночасних рівнянь є модель "попит – пропозиція", що містить функції попиту, пропозиції лінійні від ціни і умову рівноваги:

$$\begin{cases} q_t^D = \alpha_0 + \alpha_1 p_t + \varepsilon_{t1}, \alpha_1 < 0, \\ q_t^S = \beta_0 + \beta_1 p_t + \varepsilon_{t2}, \beta_1 > 0, \\ q_t^D = q_t^S, \end{cases} \quad (8.2)$$

де перше рівняння – функція попиту, друге рівняння – функція пропозиції, третє рівняння – умова рівноваги, p_t – ціна товару в момент часу, t, ε_{t1} і ε_{t2} – випадкові складові.

Змінні в системах одночасних рівнянь розподіляють на два великі класи: ендогенні змінні, значення яких визначаються всередині моделі, та екзогенні – зовнішні стосовно моделі, їх значення визначаються поза моделлю і тому вважають фіксованими. Наприклад, всі змінні в системі (8.1) є ендогенними, оскільки вони визначаються всередині системи.

Якщо розглянути найпростішу структурну форму моделі:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + \varepsilon_2, \end{cases} \quad (8.3)$$

то тут: $y-$ ендогенні змінні; $x-$ екзогенні змінні; коефіцієнти b_i і a_j є структурними коефіцієнтами моделі. Усі змінні в моделі виражені у

відхиленнях від середнього рівня, тобто за x прийнято $x - \bar{x}$, а за y прийнято $y - \bar{y}$, тому вільний член у кожному рівнянні системи відсутній.

У кейнсіанській моделі (8.2) c_t і y_t оцінюються всередині моделі (ендогенні змінні), i_t задається поза моделлю, отже, вона є екзогенною змінною. Слід зауважити, що змінні c_t і y_t можуть бути виражені через i_t і ε_t, а саме:

$$\begin{cases} y_t = \dfrac{\beta_0}{1-\beta_1} + \dfrac{1}{1-\beta_1} i_t + \dfrac{\varepsilon_t}{1-\beta}, \\ c_t = \dfrac{\beta_0}{1-\beta_1} + \dfrac{\beta_1}{1-\beta_1} i_t + \dfrac{\varepsilon_t}{1-\beta}, \end{cases} \tag{8.4}$$

де коефіцієнт $\dfrac{1}{1-\beta_1}$ є грошовий мультиплікатор, що визначає, на яку величину збільшується сукупний дохід зі збільшенням обсягу інвестицій на одиницю.

Рівняння, що складають вихідну модель, називають структурними рівняннями моделі. У свою чергу, їх підрозділяють на поведінкові рівняння (описують взаємодію між змінними) та рівняння – тотожності (співвідношення, які повинні виконуватися в усіх випадках не містять параметрів і випадкових складових).

Якщо використати МНК для оцінювання структурних коефіцієнтів моделі, то оцінки будуть зміщеними і неспроможними. Тому для визначення структурних коефіцієнтів моделі структурну форму моделі перетворюють у приведену форму моделі. Рівняння, в яких відображена схема визначення ендогенних змінних, називають рівняннями в приведеній формі (приведеними рівняннями). *Приведеними рівняннями* називають рівняння, в яких ендогенні змінні виражені тільки через екзогенні або зумовлені змінні, а також випадкові складові. *Зумовленими змінними* називаються лагові ендогенні змінні, значення яких визначені до розгляду співвідношення.

Система лінійних функцій ендогенних змінних від екзогенних є приведеною формою моделі:

$$\begin{cases} \hat{y}_1 = \delta_{11} x_1 + \delta_{12} x_2 + \ldots + \delta_{1m} x_m + \varepsilon_1, \\ \hat{y}_2 = \delta_{21} x_1 + \delta_{22} x_2 + \ldots + \delta_{2m} x_m + \varepsilon_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \hat{y}_n = \delta_{n1} x_1 + \delta_{n2} x_2 + \ldots + \delta_{nm} x_m + \varepsilon_n. \end{cases} \tag{8.5}$$

Застосувавши МНК для оцінки δ, можна оцінити значення ендогенних змінних через екзогенні, а далі оцінити значення ендогенних змінних через екзогенні. Слід відмітити, що коефіцієнти приведеної форми моделі є нелінійними функціями коефіцієнтів структурної форми моделі. Це можна показати на простому прикладі. Для структурної моделі вигляду, в якій для спрощення відсутні похибки ε:

$$\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1, \\ y_2 = b_{21} y_1 + a_{22} x_2, \end{cases} \tag{8.6}$$

приведена форма моделі має вигляд:

$$\begin{cases} y_1 = \delta_{11} x_1 + \delta_{12} x_2, \\ y_2 = \delta_{21} x_1 + \delta_{22} x_2, \end{cases} \tag{8.7}$$

коли з першого рівняння структурної моделі виражається y_2:

$$\begin{cases} y_2 = \dfrac{y_1 - a_{11} x_1}{b_{12}}, \\ y_2 = b_{21} y_1 + a_{22} x_2, \end{cases} \tag{8.8}$$

У результаті отримано:

$$\frac{y_1 - a_{11} x_1}{b_{12}} = b_{21} y_1 + a_{22} x_2 . \tag{8.9}$$

Після елементарних перетворень отримано:

$$y_1 : \; y_1 = \frac{a_{11}}{1 - b_{12} b_{21}} x_1 + \frac{a_{22} b_{12}}{1 - b_{12} b_{21}} x_2 . \tag{8.10}$$

Позначивши $\delta_{11} = \dfrac{a_{11}}{1 - b_{12} b_{21}}$, $\delta_{12} = \dfrac{a_{22} b_{12}}{1 - b_{12} b_{21}}$, отримали перше рівняння приведеної форми моделі $y_1 = \delta_{11} x_1 + \delta_{12} x_2$, в якій коефіцієнти є нелінійні співвідношення коефіцієнтів структурної форми моделі. Виконавши аналогічні перетворення отримано:

$$y_2 = \frac{a_{11} b_{21}}{1 - b_{12} b_{21}} x_1 + \frac{a_{22}}{1 - b_{12} b_{21}} x_2 \tag{8.11}$$

Через позначення $\delta_{21} = \frac{a_{11} b_{21}}{1 - b_{12} b_{21}}$ і $\delta_{22} = \frac{a_{22}}{1 - b_{12} b_{21}}$ отримано друге рівняння приведеної форми моделі:

$$y_2 = \delta_{21} x_1 + \delta_{22} x_2 . \tag{8.12}$$

Приведена форма моделі дозволяє отримати значення ендогенної змінної через значення екзогенних змінних, але в ній відсутні оцінки взаємозв'язку між ендогенними змінними.

Проблема ідентифікації. У процесі переходу від приведеної форми моделі до структурної виникає проблема ідентифікації. Під проблемою ідентифікації розуміється можливість чисельного оцінювання параметрів структурних рівнянь за оцінками коефіцієнтів приведених рівнянь.

Доцільно розглянути проблему ідентифікації для структурної моделі з двома ендогенними змінними:

$$\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1 + a_{12} x_2 + ... + a_{1m} x_m , \\ y_2 = b_{21} y_1 + a_{21} x_1 + a_{22} x_2 + ... + a_{2m} x_m . \end{cases} \tag{8.13}$$

З другого рівняння слід виразити y_1:

$$y_1 = \frac{y_2}{b_{21}} - \frac{a_{21}}{b_{21}} x_1 - \frac{a_{22}}{b_{21}} x_2 - ... - \frac{a_{2m}}{b_{21}} x_m . \tag{8.14}$$

Тоді в системі створюються два рівняння для ендогенної змінної y_1 з однаковим складом екзогенних змінних, але з різними коефіцієнтами перед ними:

$$\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1 + a_{12} x_2 + ... + a_{1m} x_m , \\ y_1 = \frac{y_2}{b_{21}} - \frac{a_{21}}{b_{21}} x_1 - \frac{a_{22}}{b_{21}} x_2 - ... - \frac{a_{2m}}{b_{21}} x_m . \end{cases} \tag{8.15}$$

Отримано два варіанта для обчислення структурних коефіцієнтів однієї моделі, це пов'язано з неповною її ідентифікацією. Структурна модель у повному вигляді має в кожному рівнянні системи n ендогенних і m екзогенних

змінних і містить $n(n-1+m)$ параметрів. Наприклад, з $n=2$ і $m=3$ повний вигляд структурної моделі буде:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + a_{13}x_3, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3, \end{cases} \quad (8.16)$$

де наявні вісім структурних коефіцієнтів.

Зведена форма моделі в повному вигляді містить nm параметрів. Для наведеного прикладу має бути шість коефіцієнтів зведеної форми моделі:

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + \delta_{13}x_3, \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + \delta_{23}x_3. \end{cases} \quad (8.17)$$

На основі шести коефіцієнтів зведеної форми моделі необхідно визначити вісім структурних коефіцієнтів, що не може призвести до єдності розв'язку. У повному вигляді структурна модель містить більшу кількість параметрів, ніж зведена форма моделі. Відповідно, $n(n-1+m)$ параметрів структурної моделі не можуть бути однозначно визначеними з nm параметрів зведеної форми моделі. Щоб отримати єдино можливий розв'язок для структурної моделі, слід передбачити, що деякі структурні коефіцієнти моделі дорівнюють нулю, оскільки окремі ознаки слабко зв'язані з ендогенною змінною. Таким чином, зменшується кількість структурних коефіцієнтів моделі. У прикладі за $a_{13}=0, a_{21}=0$ структурна модель буде мати вигляд:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3, \end{cases} \quad (8.18)$$

де кількість структурних коефіцієнтів не перевищує кількості коефіцієнтів зведеної моделі.

Є й інші рекомендації: прирівняти деякі коефіцієнти один до одного, передбачаючи, що їх вплив на ендогенну змінну однаковий. Також можна використати обмеження вигляду $b_{ij} + a_{ij} = 0$.

Вихідна система рівнянь є:

а) *ідентифікованою*, якщо за оцінками коефіцієнтів зведених рівнянь можна однозначно визначити коефіцієнти структурних рівнянь (зазвичай це вдається зробити, коли кількість рівнянь для визначення коефіцієнтів структурних рівнянь в точності дорівнює кількості цих коефіцієнтів);

б) *неідентифікованою* (недовизначеною), якщо за коефіцієнтами зведених рівнянь можна отримати кілька варіантів значень коефіцієнтів структурних рівнянь (зазвичай це відбувається коли кількість рівнянь для визначення коефіцієнтів структурних рівнянь менша кількості визначених коефіцієнтів);

в) *надідентифікованою* (перевизначеною), якщо за оцінками коефіцієнтів зведених рівнянь неможливо визначити коефіцієнти структурних рівнянь. У даному випадку система, що зв'язує коефіцієнти структурних рівнянь з коефіцієнтами зведених рівнянь, є несумісною (зазвичай у таких випадках кількість рівнянь для оцінювання коефіцієнтів структурних рівнянь більша кількості визначених коефіцієнтів).

Для визначення ідентифікованих структурних рівнянь застосовуються необхідні та достатні умови. Нехай система одночасних рівнянь включає N рівнянь відносно N ендогенних змінних. Система містить M екзогенних або зумовлених змінних. Нехай n і m – кількість, відповідно, ендогенних і екзогенних змінних в рівнянні, яке перевіряється на ідентифікованість. Змінні, що не входять в дане рівняння, але входять в інші рівняння системи, називають *виключеними змінними* (їх кількість дорівнює $N-n$ для ендогенних і $M-m$ для екзогенних змінних, відповідно).

Перша необхідна умова ідентифікованості. Рівняння ідентифіковане, якщо воно виключає принаймні $N-1$ змінну (ендогенну або екзогенну), присутню в моделі:

$$(N-n)+(M-m)\geq N-1. \tag{8.19}$$

Друга необхідна умова ідентифікованості. Рівняння ідентифіковане, якщо кількість виключених з рівняння екзогенних змінних не менше кількості ендогенних змінних у цьому рівнянні, зменшеного на одиницю:

$$M-m\geq n-1. \tag{8.20}$$

Знаки рівностей в обох необхідних умовах відповідають точній ідентифікації рівнянь.

Необхідна і достатня умова ідентифікованості. У моделі, що містить N рівнянь відносно N ендогенних змінних, умова ідентифікованості виконується тоді і тільки тоді, коли ранг матриці, складеної з виключених з даного рівняння змінних, але які входять в інші рівняння системи, дорівнює $N-1$.

Методи оцінювання параметрів систем рівнянь

Найбільшого поширення набули такі методи оцінювання коефіцієнтів структурної моделі: непрямий метод найменших квадратів; двокроковий метод найменших квадратів; трьохкроковий метод найменших квадратів; метод максимальної правдоподібності з повною інформацією; метод максимальної правдоподібності при обмеженій інформації.

Непрямий метод найменших квадратів (НМНК). Безпосереднє використання МНК для оцінювання параметрів кожного з рівнянь системи одночасних рівнянь призводить до отримання зміщених і неспроможних оцінок. Як правило, це відбувається внаслідок корельованості однієї або декількох пояснювальних змінних з випадковим відхиленням. Тому застосовують інші методи.

Непрямий метод найменших квадратів (НМНК) заснований на використанні приведених рівнянь. Така назва методу обумовлена обчисленням оцінок b_1 і b_2 через оцінки приведених рівнянь. Оцінки b_1 і b_2, які одержані з НМНК, є спроможними. Слід також зазначити, що оцінки b_1 і b_2 визначаються однозначно, а відповідне рівняння називають ідентифікованим.

Непрямий метод найменших квадратів включає в себе такі етапи:

1) виходячи із структурних рівнянь, будуються рівняння в приведеній формі;

2) оцінюють за МНК параметри рівнянь в приведеній формі;

3) на основі знайдених оцінок, оцінюються параметри структурних рівнянь.

Двокроковий метод найменших квадратів. У системах одночасних рівнянь двохкроковий МНК застосовується для оцінки параметрів структурних рівнянь, оскільки в останніх в якості факторів беруть участь ендогенні змінні моделі і застосування звичайного МНК призводить до зміщених і неможливих оцінок.

Тут в якості додаткових змінних зазвичай виступають екзогенні змінні самої моделі. Відповідно, процедура оцінювання полягає в тому, що на першому кроці звичайним МНК оцінюється регресія ендогенних змінних на всі екзогенні змінні системи, а потім ці оцінки використовують на другому кроці замість ендогенних змінних правій частині структурного рівняння, до якого застосовується звичайний МНК.

Такий підхід дозволяє отримати можливі оцінки параметрів структурної форми.

Непрямий і двокроковий методи найменших квадратів розглядаються як традиційні методи оцінювання коефіцієнтів структурної моделі. Ці методи не складно реалізувати. Непрямий метод найменших квадратів застосовується для ідентифікованої системи одночасних рівнянь, а двокроковий метод найменших квадратів використовується для оцінювання коефіцієнтів надідентифікованої моделі.

Метод максимальної правдоподібності розглядається як найбільш загальний метод оцінювання, результати якого за нормальним розподілом ознак збігаються з МНК. Однак за великого числа рівнянь системи цей метод призводить до досить складних обчислювальних процедур. Тому як модифікація служить метод *максимальної правдоподібності з обмеженою інформацією* (метод найменшого дисперсійного відношення). Однак не зважаючи на популярність даного методу в 60-х роках XIX ст., він був практично витіснений двокроковим методом найменших квадратів (ДМНК) у зв'язку з більшою простотою останнього.

Подальшим розвитком ДМНК є *трикроковий метод найменших квадратів (ТМНК)*. Цей метод оцінювання придатний для всіх видів рівнянь структурної моделі.

Для практичного застосування ЗМНК потрібно виконати такі вимоги:

1) усі тотожності, які входять до системи рівнянь, виключають з розгляду, тому що вони не містять невідомих параметрів і не параметризуються;

2) кожне неідентифіковане рівняння також виключають із системи, оскільки оцінити їх параметри в принципі неможливо;

3) точно ідентифіковані та надідентифіковані рівняння поділяють на дві різні групи і ЗМНК застосовують до кожної з них окремо;

4) якщо група надідентифікованих рівнянь складається лише з одного рівняння, то ЗМНК перетворюється на 2МНК;

5) кореляція залишків окремих рівнянь системи призводить до того, що загальна матриця коваріацій системи є недіагональною, однак водночас не між усіма рівняннями системи існує залежність, тому матриця коваріації часто буває блочно-діагональною, тоді оцінювання параметрів на основі ТМНК виконують окремо для кожної групи рівнянь, що відповідають одному блоку.

Однак через деякі обмеження на параметри ефективнішим є ДМНК.

8.2. Теоретичні відомості про функції *MATLAB*, які використовуються в даній лабораторній роботі

Symbolic Math Toolbox дозволяє користуватися символьною математикою й обчисленнями із плаваючою точкою в *MATLAB*. Пакет включає обчислювальне ядро пакета Maple V, розробленого фірмою *Waterloo Maple Software.*

Extended Symbolic Math Toolbox надає користувачеві додаткову можливість програмування на Maple і забезпечує доступ до спеціалізованих бібліотек Maple.

`solve` **Розв'язання рівнянь у символьному виді**

Синтаксис

```
solve(E1, E2, ..., EN)
solve(E1, E2, ..., EN, var1, var2, ..., var)
```

Опис

Тут *E*1, *E*2, ..., *EN* – символьні вирази або змінні, у яких вони містяться, а *var*1, ..., *varN* – змінні, відносно яких розв'язується система рівнянь *E*1=0, *E*2=0, ..., *EN*=0. Зрозуміло, перша форма виклику цієї функції допустима лише у випадку, коли немає неоднозначностей щодо того, що саме варто знайти. Функція *solve* повертає єдине символьне вираження, якщо рівняння (система рівнянь) має єдиний розв'язок і вектор розв'язків у протилежному випадку. Якщо рівняння містить періодичні функції і тому може мати нескінченне число розв'язків, функція обмежується тим, що повертає розв'язок за один період в околі нуля.

`DET` **Визначник матриці**

Синтаксис

```
d = det(A)
```

Опис

Функція *d = det*(*A*) обчислює визначник квадратної матриці; якщо матриця *A* цілочисельна, то результатом є також ціле число.

Алгоритм

Визначник матриці обчислюється на основі трикутного розкладання методом виключення Гаусса:

```
[L, U] = lu(A);
s = det(L);
d = s * prod(diag(U)).
```

RANK **Ранг матриці**

Синтаксис

```
r = rank(A)
r = rank(A, tol)
```

Опис

Функція `r = rank(A)` повертає ранг матриці, що визначається як кількість сингулярних чисел, що перевищують поріг `max(size(A))*norm(A)*eps`.

Функція `r = rank(A, tol)` повертає ранг матриці, що визначається як кількість сингулярних чисел, що перевищують заданий поріг `tol`.

Алгоритм

Існує кілька підходів до обчислення рангу матриці. У системі *MATLAB* використаний метод, заснований на обчисленні сингулярних чисел матриці `A`; він реалізований у вигляді функції `svd`. Це найбільш надійний метод, хоча й потребує значного часу на обчислення.

Сам алгоритм обчислення рангу досить простий:

```
s = svd(A);
tol = max(size(A)) * s(1) * eps;
r = sum(s > tol).
```

8.3. Розв'язування типової задачі в середовищі *MATLAB*

Задача. Побудувати модель залежностей експорту і імпорту країни y_1, y_2 (млн дол.) від ВВП x_1 (млн дол.) і ціни на нафту x_2 (дол./бар.) на підставі даних, наведених у табл.8.1.

Таблиця 8.1

Вихідні дані задачі

Рік	ВВП, млн. дол., x_1	Ціна нафти, дол. /бар., x_2	Експорт, млн. дол., y_1	Імпорт, млн. дол., y_2
1	2	3	4	5
1998	2 630	11,9	74,4	58,0
1991	4823	17,1	75,6	39,5

Закінчення табл. 8.1

1	2	3	4	5
2000	7 306	26,7	105,0	44,9
2001	8 944	23,0	101,9	53,8
2002	10 819	23,9	107,3	61,0
2003	13 208	27,3	135,9	76,1
2004	17 027	34,2	183,2	97,4
2005	21 610	50,0	243,8	125,4
2006	26 917	61,1	303,6	164,3
2007	33 248	68,9	354,4	223,5
2008	41 277	94,8	471,6	291,9

Для початку роботи необхідно створити новий М-файл. Для цього з меню File вибрати опцію New, а потім M-File.

У вікні, що з'явилося, редактора М-файлів ввести вихідні дані.

Фрагмент М-файла

```
%% Системи одночасних рівняньclc
clear all
y1 = [74.4 75.6 105.0 101.9 107.3 135.9 183.2 243.8 303.6 354.4
471.6]';
y2 = [58.0 39.5 44.9 53.8 61.0 76.1 97.4 125.4 164.3 223.5
291.9]';
x1 = [2630 4823 7306 8944 10819 13208 17027 21610 26917 33248
41277]';
x2 = [11.9 17.1 26.7 23.0 23.9 27.3 34.2 50.0 61.1 68.9 94.8]';
[n1, m1] = size(x1);
[n2, m2] = size(x2);
```

З урахуванням даних, наведених у табл. 8.1, система одночасних рівнянь має вигляд:

$$\begin{cases} y_1 = b_{12} y_2 + a_{01} + a_{11} x_1 + \varepsilon_1 \\ y_2 = b_{21} y_1 + a_{02} + a_{22} x_2 + \varepsilon_2 \end{cases}. \qquad (8.21)$$

Ця економічна модель показує, що обсяг експорту прямо пропорційний ВВП країни й обсягу імпорту. У свою чергу, на імпорт країни впливає ціна нафти й обсяги експорту. Зрозуміло, що зі зростанням експорту повинен збільшуватися імпорт. Таким чином, отримано систему взаємозалежних (спільних) рівнянь, модель яких описується формулою (8.3).

Ідентифікацію кожного рівняння моделі можна виконати, скориставшись співвідношенням (8.5) і достатньою умовою ідентифікації.

Фрагмент М-файла

```
%% Ідентифікація моделі
% Перше рівняння:
H = 2;% Ендогенні змінні
D = 1;% Екзогенні змінні
if D+1 == H
    fprintf('\nH:рівняння ідентифіковане \n')
end
if D+1 < H
    fprintf('\nH:рівняння не є ідентифікованим\n')
end
if D+1 > H
    fprintf('\nH: рівняння надідентифіковане\n')
end
syms a22 % коефіцієнти при змінних,
         % що відсутні в досліджуваному рівнянні
A = [a22]; % Матриця, складена з
           % цих коефіцієнтів
if (det(A) ~= 0) && (double(rank(A)) >= H-1)
     fprintf('\nД: рівняння ідентифіковане \n')
end
```

```
>> Перше рівняння

H:рівняння ідентифіковане

Д: рівняння ідентифіковане

>> Друге рівняння

H: рівняння ідентифіковане

Д: рівняння ідентифіковане
```

Таким чином, система одночасних рівнянь ідентифікована в силу ідентифікованості рівнянь (як першого, так і другого). Для оцінювання параметрів системи можна застосовувати як непрямий, так і двокроковий МНК.

Непрямим методом найменших квадратів оцінюють параметри структурної моделі:

1) складемо приведену форму моделі:

$$\begin{cases} \hat{y}_1 = \delta_{10} + \delta_{11}x_1 + \delta_{12}x_2 \\ \hat{y}_2 = \delta_{20} + \delta_{21}x_1 + \delta_{22}x_2 \end{cases}. \qquad (8.22)$$

2) для кожного рівняння приведеної форми стандартним МНК визначити його коефіцієнти:

Фрагмент М-файла

```
%% Непрямий МНК
fprintf('\n Приведена форма моделі має вигляд:\n')
X = [ones(n1,1) x1 x2];
[b1,bint1,r1,rint1,stats1] = regress(y1,X,0.05);
[b2,bint2,r2,rint2,stats2] = regress(y2,X,0.05);
yp1 = b1(1) + b1(2)*x1 + b1(3)*x2;% перше рівняння структурної форми моделі
yp2 = b2(1) + b2(2)*x1 + b2(3)*x2;% друге рівняння структурної форми моделі
fprintf('\nПерше рівняння\n')
fprintf('\n yp1 = %f + %f*x1 + %f *x2',b1)
R21 = stats1(1)% коефіцієнт детермінації
F1 = stats1(2)%F-статистика
Ft1 = finv(0.95,m1,n1-m1-1);
if F1 > Ft1
    fprintf('\nРівняння значуще в цілому\n')
else
    fprintf('\nРівняння не є значищим\n')
end
fprintf('\nДруге рівняння\n')
fprintf('\n yp2 = %f + %f*x1 + %f *x2\n',b2)
R22 = stats2(1)% коефіцієнт детермінації
F2 = stats2(2)%F-статистика
Ft2 = finv(0.95,m2,n2-m2-1);
if F2 > Ft2
    fprintf('\nРівняння значуще в цілому\n')
else
    fprintf('\nРівняння не є значущим\n')
end
```

```
>> Приведена форма моделі має вигляд:

>> Перше рівняння

 yp1 = 0.158201 + 0.003715*x1 + 3.320324 *x2
```

```
R21 =

    0.9940

F1 =

  663.8553

Рівняння значуще в цілому

>> Друге рівняння

 yp2 = -7.109337 + 0.002903*x1 + 1.751495 *x2

R22 =

    0.9624

F2 =

  102.5016

Рівняння значуще в цілому
```

Таким чином, приведена форма моделі має вигляд:

$$\begin{cases} \hat{y}_1 = 0.1582 + 0.0037x_1 + 3.3203x_2 + \varepsilon_1' \\ \hat{y}_2 = -7.1093 + 0.0029x_1 + 1.7515x_2 + \varepsilon_2' \end{cases}; \qquad (8.23)$$

3) перейти від приведеної до структурної форми моделі, використовуючи функцію *solve* для символьних обчислень в *MATLAB*.

Фрагмент М-файла

```
syms y11 y22 x11 x22;
x11 = solve('y11 = 0.158201 + 0.003715*x11 + 3.320324 *x22',x11)
x22 = solve('y22 = -7.109337 + 0.002903*x11 + 1.751495 *x22',x22)
```

```
>> x11 = 269.1790*y11 - 893.7615*x22 - 42.5844
>> x22 = 0.5709*y22 - 0.0017*x11 + 4.0590
```

Отже, структурна форма моделі має вигляд:

$$\begin{cases} \hat{y}_1 = 1{,}8956y_2 + 13.6353 - 0.0019x_1 + \varepsilon_1 \\ \hat{y}_2 = 0.7806y_1 - 7.2328 - 0.8404x_2 + \varepsilon_2 \end{cases}. \qquad (8.24)$$

Оцінити кожне рівняння моделі можна, використовуючи коефіцієнт детермінації й *F*-критерій Фішера.

Фрагмент М-файла

```
fprintf('\nКритерій якості\n')
fprintf('\nПерше рівняння:\n')
n1 = 11;
m1 = 2;
y1p = 1.8956*y2 + 13.653 - 0.0019 * x1;
R2 = 1 - sum((y1 - y1p).^2)/sum((y1 - mean(y1)).^2)
R2_adjasted = 1 - ((n1 - 1)/(n1 - m1 - 1))*(1-R2)
F1p = (sum((y1p - mean(y1)).^2)*(n1 - m1))/(sum((y1-y1p).^2)*(m1 -
1))
F1p = (R2*(n1 - m1 - 1))/((1 - R2)*m1)
F1t = finv(0.95,m1,n1-m1-1)
if F1p > F1t
    fprintf('\n Рівняння значуще в цілому\n')
else
    fprintf('\n Рівняння не є значущим\n')
end
```

```
>> Критерій якості
Перше рівняння:
R2 =
    0.9689
R2_adjasted =
    0.9612
F1p =
  297.4983
F1t =
    4.4590
Рівняння значуще в цілому
```

Фрагмент М-файла

```
fprintf('\nДруге рівняння:\n')
n2 = 11;
m2 = 2;
```

```
y2p = 0.7806*y1 - 7.2328 - 0.8404 * x2;
R2 = 1 - sum((y2 - y2p).^2)/sum((y2 - mean(y2)).^2)
R2_adjasted = 1 - ((n2 - 1)/(n2 - m2 - 1))*(1-R2)
F2p = (sum((y2p - mean(y2)).^2)*(n2 - m2))/(sum((y2-y2p).^2)*(m2 -
1))
F2p = (R2*(n2 - m2 - 1))/((1 - R2)*m2)
F2t = finv(0.95,m2,n2-m2-1)
if F2p > F1t
    fprintf('\n Рівняння значуще в цілому\n')
else
    fprintf('\n Рівняння не є значущим\n')
end
```

```
Друге рівняння:

R2 =

    0.9828

R2_adjasted =

    0.9785
F2p =

  506.9962
F2t =

    4.4590

Рівняння значуще в цілому
```

Дана система показує, що для *першого рівняння* ВВП країни (екзогенна змінна x_1) практично незначущо впливає на обсяг експорту, на який впливає обсяг імпорту (ендогенна змінна y_2). Коефіцієнт детермінації $R^2 = 0.9612$ показує, що включені в модель змінні (x_1 й y_2) пояснюють 96,12 % змінюваності змінної y_1, а інша її змінюваність пояснюється неврахованими в моделі факторами. Виходячи з обчисленого та табличного значень *F*-критерію Фішера можна стверджувати, що отримане рівняння є значущим у цілому.

Для *другого рівняння* на обсяг імпорту впливає обсяг експорту (ендогенна змінна y_1). Помітно, що між ціною на нафту й обсягами імпорту існує обернено пропорційний зв'язок: чим більша ціна на нафту (екзогенна змінна x_2), тим менший обсяг імпорту. Коефіцієнт детермінації $R^2 = 0.9785$ показує, що включені в модель змінні (x_2 і y_1) пояснюють 97,85 % змінюваності змінної y_2,

а інша її змінюваності пояснюється неврахованими в моделі причинами. Виходячи з обчисленого та табличного значень *F*-критерію Фішера можна стверджувати, що отримане рівняння є значущим у цілому.

До недоліків даної моделі можна віднести те, що не всі коефіцієнти моделі є значущими (перевірку здійснюють з використанням критерію Стьюдента):

Фрагмент М-файла

```
D = [ones(n1,1) y2 x1];
C = inv(D'*D);
tt = tinv(0.95,n1-m1-1);
tb1 = 1.8956/(sqrt(sum((y1 - y1p).^2) / (n1 - m1 -
1))*sqrt(C(2,2)))
if tb1 > tt
    fprintf('\nПараметр за y2 є значущим\n')
else
    fprintf('\nПараметр за y2 не є значущим\n')
end
tb2 = 0.0019/(sqrt(sum((y1 - y1p).^2) / (n1 - m1 -
1))*sqrt(C(3,3)))
if tb2 > tt
    fprintf('\nПараметр за x1 є значущим\n')
else
    fprintf('\nПараметр за x1 не є значущим\n')
end
D = [ones(n2,1) y1 x2];
C = inv(D'*D);
tt = tinv(0.95,n2-m2-1);
tb1 = 0.7806/(sqrt(sum((y2 - y2p).^2) / (n2 - m2 -
1))*sqrt(C(2,2)))
if tb1 > tt
    fprintf('\nПараметр за y1 є значущим\n')
else
    fprintf('\nПараметр за y1 не є значущим\n')
end
tb2 = 0.8404/(sqrt(sum((y2 - y2p).^2) / (n2 - m2 -
1))*sqrt(C(3,3)))
if tb2 > tt
    fprintf('\nПараметр за x2 є значущим\n')
else
    fprintf('\nПараметр за x2 не є значущим\n')
end
```

```
tb1 = 4.0435
```

```
Параметр за y2 є значущим

tb2 = 0.6103

Параметр за x1 не є значущим
tb1 = 2.5540

Параметр за y1 є значущим

tb2 = 0.5378

Параметр за x2 не є значущим
```

Для реалізації двокрокового методу найменших квадратів необхідно побудувати надідентифіковану модель. Для цього слід прирівняти коефіцієнти $b_{12} = a_{11}$:

$$\begin{cases} y_1 = a_{01} + b_{12}(y_2 + x_1) + \varepsilon_1 \\ y_2 = b_{21}y_1 + a_{02} + a_{22}x_2 + \varepsilon_2 \end{cases}. \qquad (8.25)$$

Виконати ідентифікацію кожного рівняння моделі можна скориставшись співвідношенням (8.5) і достатньою умовою ідентифікації.

Фрагмент М-файла

```
% Ідентифікація моделі:
% Перше рівняння:
H = 1;% Ендогенні змінні
D = 1;% Екзогенні змінні
if D+1 == H
    fprintf('\nH:рівняння ідентифіковане\n')
end
if D+1 < H
    fprintf('\nH:рівняння не є ідентифікованим\n')
end
if D+1 > H
    fprintf('\nH:рівняння надідентифіковане\n')
end
A = [0]; % Матриця, складена з
         % цих коефіцієнтів при змінних
if (det(A) ~= 0) && (double(rank(A)) >= H-1)
     fprintf('\nД:рівняння ідентифіковане\n')
else
    fprintf('\nД:рівняння не є ідентифікованим\n')
end
```

```
Н: рівняння надідентифіковане

Д: рівняння не є ідентифікованим

Н: рівняння ідентифіковане

Д: рівняння ідентифіковане
```

Фрагмент М-файла

```
% Друге:
H = 2;% Ендогенні змінні
D = 1;% Екзогенні змінні
if D + 1 == H
     fprintf ('\ nH: рівняння ідентифіковане \ n')
end
if D + 1 <H
     fprintf ('\ nH: рівняння не є ідентифікованим \ n')
end
if D + 1> H
     fprintf ('\ nH: рівняння надідентифіковане \ n')
end
syms a11% коефіцієнти при змінних,
        % відсутніх в досліджуваному рівнянні
A = [a11]; % Матриця, складена з
           % цих коефіцієнтів при змінних
if (det (A) ~ = 0) && (double (rank (A))> = H-1)
      fprintf ('\ НД: рівняння ідентифіковане \ n')
else
      fprintf ('\ НД: рівняння не є ідентифікованим\ n')
end
```

Для визначення параметрів надідентифікованої моделі використовують двокроковий метод найменших квадратів.

Крок 1. На першому кроці треба знайти приведену форму моделі:

$$\begin{cases} \hat{y}_1 = \delta_{10} + \delta_{11}x_1 + \delta_{12}x_2 \\ \hat{y}_2 = \delta_{20} + \delta_{21}x_1 + \delta_{22}x_2 \end{cases}. \tag{8.26}$$

Припускаючи використання тих же вихідних даних, що й у попередньому прикладі, отримують ту ж систему приведених рівнянь:

$$\begin{cases} \hat{y}_1 = 0.1582 + 0.0037x_1 + 3.3203x_2 + \varepsilon_1' \\ \hat{y}_2 = -7.1093 + 0.0029x_1 + 1.7515x_2 + \varepsilon_2' \end{cases} \tag{8.27}$$

Крок 2. На основі системи приведених рівнянь за точно ідентифікованим другим рівнянням можна знайти теоретичні значення для ендогенної змінної y_2. Із цією метою в друге приведене рівняння підставити значення x_1 і x_2.

Крок 3. За надідентифікованим рівнянням структурної форми моделі замінити фактичне значення y_2 на теоретичні y_{2p} і розрахувати нову змінну $z = y_{2p} + x_1$:

Фрагмент М-файла

```
printf ('\ Приведена форма моделі має вигляд: \ n')
X = [ones (n1,1) x1 x2];
[b1, bint1, r1, rint1, stats1] = regress (y1, X, 0.05);
[b2, bint2, r2, rint2, stats2] = regress (y2, X, 0.05);
yp1 = b1 (1) + b1 (2) * x1 + b1 (3) * x2;% перше рівняння
структурної форми моделі
fprintf ('\ nПерше рівняння \ n')
fprintf ('\ n yp1 =% f +% f * x1 +% f * x2', b1)
yp2 = b2 (1) + b2 (2) * x1 + b2 (3) * x2;% друге рівняння
структурної форми моделі
fprintf ('\ nДруге рівняння \ n')
fprintf ('\ n yp2 =% f +% f * x1 +% f * x2 \ n', b2)
% Розрахункові дані для третього кроку ДМНК
z = yp2 + x1;
[x1 yp2 z]
```

```
>> Приведена форма моделі має вигляд:

>> Перше рівняння

 yp1 = 0.158201 + 0.003715*x1 + 3.320324 *x2

>> Друге рівняння

 yp2 = -7.109337 + 0.002903*x1 + 1.751495 *x2

ans =

  1.0e+004 *

        0.2630    0.0021    0.2651
        0.4823    0.0037    0.4860
        0.7306    0.0061    0.7367
        0.8944    0.0059    0.9003
        1.0819    0.0066    1.0885
```

```
1.3208    0.0079    1.3287
1.7027    0.0102    1.7129
2.1610    0.0143    2.1753
2.6917    0.0178    2.7095
3.3248    0.0210    3.3458
4.1277    0.0279    4.1556
```

Крок 4. Далі до надідентифікованого рівняння застосовується метод найменших квадратів: $y_1 = a_{01} + b_{12} z + \varepsilon_1$.

Фрагмент М-файла

```
% МНК для надідентифікованого рівняння:
X = [ones (n1,1) z];
[b3, bint3, r3, rint3, stats3] = regress (y1, X, 0.05);
fprintf ('\ nПерше рівняння структурної форми має вигляд: \ n')
y1p = b3 (1) + b3 (2) * z;
fprintf ('\ n y1p =% f +% f * z \ n', b3)
```

```
>> Перше рівняння структурної форми має вигляд:

 y1p = 16.539257 + 0.010446*z

>> Друге рівняння

 y2p = 0.7806*y1 - 7.2328 - 0.8404 * x2
```

У цілому розглянута система одночасних рівнянь складе:

$$\begin{cases} \hat{y}_1 = 16.5393 + 0.0104(y_2 + x_1) + \varepsilon_1 \\ \hat{y}_2 = -7.2328 + 0.7806 y_1 - 0.8404 x_2 + \varepsilon_2 \end{cases}. \qquad (8.28)$$

Таким чином, двокроковий метод найменших квадратів є найбільш загальним і широко розповсюдженим методом розв'язання системи одночасних рівнянь. Для точно ідентифікованих рівнянь ДМНК дає той же результат, що й КМНК.

Запитання для самоперевірки

1. Які основні причини використання систем одночасних рівнянь?
2. Наведіть основні типи моделей застосування систем одночасних рівнянь.
3. Які види змінних розрізняють в системах одночасних рівнянь? Дайте їм визначення.

4. У чому полягає відмінність між структурними рівняннями системи та рівняннями у приведеній формі?

5. Чому звичайний МНК практично не використовується для оцінювання систем одночасних рівнянь?

6. Які найбільш поширені методи оцінювання коефіцієнтів структурної моделі?

7. У чому полягає сутність непрямого методу найменших квадратів?

8. Назвіть причини неідентифікованості та надідентифікованості систем одночасних рівнянь.

9. Наведіть необхідні та достатні умови ідентифікованості систем.

Завдання для лабораторної роботи

Задача. Побудувати модель споживання свинини на душу населення y_1 (у фунтах) залежно від ціни на неї y_2 (дол./фунт), передбачуваного доходу x_1 (у дол.) і витрат на обробку м'яса x_2 (% від ціни). Статистичні дані для даної задачі наведені в табл. 8.2 ($p_1 = p_{1/10}$ – кількість букв у повнім імені, $p_2 = p_{2/10}$ – – кількість букв у прізвищі).

Таблиця 8.2

Вхідні дані задачі

№ п/п	y_1	y_2	x_1	x_2
1	2, 914	-0,12	$0,1 + p_1$	$7, 452 – p_2$
2	0, 925	0, 098	$0,2 + p_1$	$2, 196 – p_2$
3	1, 976	0, 108	$0,3 + p_1$	$4,83 – p_2$
4	5, 386	-0,05	$0,4 + p_1$	$13,56 – p_2$
5	2, 683	0, 209	$0,5 + p_1$	$6, 451 – p_2$
6	4, 779	0, 155	$0,6 + p_1$	$11,79 – p_2$
7	5, 413	0, 186	$0,7 + p_1$	$13,35 – p_2$
8	3, 973	0, 353	$0,8 + p_1$	$9, 497 – p_2$
9	5, 015	0, 367	$0,9 + p_1$	$12,1 – p_2$
10	7, 807	0, 262	$1,0 + p_1$	$19,25 – p_2$
11	6, 371	0, 424	$1,1 + p_1$	$15,41 – p_2$

1. За вихідним даними скласти структурну форму моделі.

2. Перевірити модель на ідентифікацію.

3. Непрямим методом найменших квадратів оцінити параметри структурної моделі.

3.1. Скласти приведену форму моделі.

3.2. Для кожного рівняння приведеної форми стандартним МНК визначити його коефіцієнти.

3.3. Перейти від приведеної до структурної форми моделі.

3.4. Оцінити кожне рівняння моделі, використовуючи *F*-критерій та коефіцієнт детермінації. Дати економічну інтерпретацію отриманих результатів.

4. Побудувати надідентифіковану модель, виходячи з попередньої моделі.

5. Двокроковим методом найменших квадратів оцінити параметри моделі.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

Лабораторна робота 9
Аналіз часових рядів

Мета роботи: навчитися будувати моделі часових рядів, виділяючи тренд і сезонні коливання; за побудованими моделями вміти здійснювати точковий прогноз і графічно реалізовувати побудовану модель.

Основні задачі лабораторної роботи:

1. За вихідним даними зробити згладжування ряду (усунути циклічні коливання з часового ряду).

2. Визначити вид моделі часового ряду (адитивна або мультиплікативна).

3. Виділити й усунути сезонні коливання з часового ряду.

4. Визначити вид функції тренда.

5. Оцінити параметри тренда й усунути його з часового ряду.

6. Використовуючи вбудовану функцію `autocorr` розрахувати коефіцієнти кореляції та побудувати корелограму.

7. Здійснити прогноз обсягу товарообігу на перший квартал 1997 року.

8. Дати графічну інтерпретацію побудованої моделі часового ряду.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

9.1. Основні поняття часових рядів

В аналізі багатьох економічних показників використовують щорічні, щоквартальні, щомісячні, щоденні дані. Прикладом цього є річні дані про валовий національний продукт (ВНП), випуск валової продукції (ВВП), обсяг чистого експорту, місячні дані про обсяг продажу продукції, щоденні дані щодо виробництва продукції на конкретному підприємстві. Для об'єктивності аналізу дані

слід систематизувати в часі. Упорядковані дані за часом їх отримання є **часовим рядом.**

Моделі, побудовані на основі даних, що характеризують один об'єкт за рядом послідовних моментів (періодів) часу, називаються **моделями часових рядів.**

Кожен рівень часового ряду формується під впливом великої кількості факторів, які умовно можна поділити на три групи:

1) фактори, що формують тенденцію ряду;

2) фактори, що формують циклічні коливання ряду;

3) випадкові фактори.

Кожен рівень часового ряду формується під впливом тенденції, сезонних коливань і випадкової компоненти. Процедура ідентифікації називається **декомпозицією**. Кожну компоненту ідентифікують окремо. Потім вклади кожної компоненти комбінують з метою отримання прогнозів щодо майбутніх значень часових рядів. Методи декомпозиції використовують як для короткочасних, так і для довготривалих прогнозів. З їх допомогою можна відображувати зростання або спадання, що є основою ряду, або коригувати значення ряду, виключаючи з них одну або кілька компонент.

Слід зазначити, що останнім часом до прогнозів, зроблених на основі методу декомпозиції, ставляться як до проміжних, а сам метод розглядають як інструмент розуміння часових рядів.

Метод декомпозиції припускає виділення компонент: трендової, циклічної, сезонної та випадкової. **Тренд** – це компонента, що представляє основне зростання (спадання) у часовому ряді. Трендова компонента утворюється під впливом постійної зміни факторів, її позначають буквою *T*. **Циклічна компонента** − послідовність хвилеподібних флуктуацій або цикли більше одного року. Зміна економічних умов зазвичай відбувається циклічно. Позначається циклічна компонента буквою *C*. На практиці ідентифікувати цикл складно, він часто здається частиною тренда. У цьому випадку компоненту називають *трендово-циклічною* і позначають буквою *T*. **Сезонні зміни** зазвичай присутні в квартальних, місячних і тижневих даних. Під сезонними варіаціями розуміються зміни з певною стабільною структурою, що мають річну циклічність. Сезонну компоненту позначають буквою *S*. **Випадкова компонента** обумовлена впливом безлічі різноманітних подій, які як такі несуттєві, але сукупно можуть дати значний ефект *E*. У різних поєднаннях у досліджуваному явищі або процесі ці компоненти можуть приймати різні форми.

У більшості випадків фактичний рівень часового ряду можна подати як суму або добуток трендової, циклічної та випадкової компонент. Двома

найпростішими моделями, які зв'язують величину часового ряду, що спостерігається y_t з компонентами тренду T, сезонності S і випадковості E, є моделі адитивних і мультиплікативних компонент. Модель, в якій часовий ряд поданий як сума перерахованих компонент, називають *адитивною моделлю часового ряду*: $Y_t = T_t + S_t + E_t$. Модель, в якій часовий ряд поданий як добуток перерахованих компонент, називається *мультиплікативною моделлю часового ряду*: $Y_t = T_t \cdot S_t \cdot E_t$. Модель адитивних компонент застосовують коли аналізований часовий ряд має приблизно однакові зміни протягом всієї тривалості ряду, тобто всі значення ряду істотно зменшуються в межах смуги постійної ширини, центрованої на рівні ряду. Модель мультиплікативних компонент ефективніша в тих випадках, коли зміна часової послідовності збільшується зі зростанням рівня, тобто значення ряду розходяться як такі, що мають тренд, а послідовність спостережуваних значень нагадує рупор або лійку.

Основне завдання економетричного дослідження окремого часового ряду – виявлення та надання кількісного вираження кожній із зазначених компонент з тим, щоб використовувати отриману інформацію для прогнозування майбутніх значень ряду або у побудові моделей взаємозв'язку двох або більше часових рядів.

За наявності у часовому ряді тенденції та циклічних коливань значення кожного наступного рівня ряду залежать від попередніх. Кореляційну залежність між послідовними рівнями часового ряду називають *автокореляцією рівнів ряду*. Кількісно її можна виміряти за допомогою лінійного коефіцієнта кореляції між рівнями вихідного часового ряду та рівнями цього ряду, зсунутими на кілька кроків у часі.

$$r_1 = \frac{\sum_{t=2}^{n} (y_t - \overline{y_1})(y_{t-1} - \overline{y_2})}{\sqrt{\sum_{t=2}^{n} (y_t - \overline{y_1})^2 (y_{t-1} - \overline{y_2})^2}}. \qquad (9.1)$$

Аналогічно можна визначити коефіцієнти автокореляції другого і більш високих порядків (9.2).

$$r_2 = \frac{\sum_{t=2}^{n} (y_t - \overline{y_3})(y_{t-2} - \overline{y_4})}{\sqrt{\sum_{t=2}^{n} (y_t - \overline{y_3})^2 (y_{t-2} - \overline{y_4})^2}}. \qquad (9.2)$$

Кількість періодів, за якими розраховується коефіцієнт автокореляції, називається *лагом*. Є думка, що максимальний лаг повинен бути не більшим за $\frac{n}{4}$.

Коефіцієнт автокореляції характеризує тісноту тільки лінійного зв'язку поточного та попереднього рівнів ряду. Для деяких часових рядів, що мають сильну нелінійну тенденцію, коефіцієнт автокореляції може наближатися до нуля. За знаком коефіцієнта автокореляції не можна робити висновок про зростаючу або спадну тенденції в рівнях ряду.

Послідовність коефіцієнтів автокореляції рівнів першого, другого і так далі порядків називають *автокореляційною функцією часового ряду*. Графік залежності її значень від величини лага (порядку коефіцієнта автокореляції) називається *корелограмою*. За допомогою аналізу автокореляційної функції і корелограми можна виявити структуру ряду.

Під час аналізу часового ряду існує основне завдання – визначення основної тенденції у розвитку досліджуваного явища. У деяких випадках загальна тенденція простежується в динаміці показника, в інших ситуаціях вона не може проглядатися через існуючі випадкові коливання. Наприклад, в окремі моменти часу сильні коливання в курсах акцій можуть заступити наявність тенденції до зростання або зниження цього показника. На практиці простим методом виявлення загальної тенденції є укрупнення інтервалів. Наприклад, ряд тижневих даних можна перетворити на ряд місячних даних, ряд квартальних даних − на річні. Таким перетворенням може бути підсумовування рівнів вихідного ряду або знаходження середніх значень. Цей метод називається *згладжуванням часового ряду*. Сутність цього методу полягає в заміні фактичних рівнів часового ряду розрахунковими, які менш схильні до коливань і сприяють більш чіткому прояву тенденції розвитку.

Згідно з різними підходами методи згладжування поділяють на дві групи: аналітичний та алгоритмічний. *Аналітичний підхід* припускає завдання загального вигляду функції, що описує регулярну, невипадкову складову. Надалі проводять статистичне оцінювання невідомих коефіцієнтів моделі і визначають згладжені значення рівнів часового ряду шляхом підстановки відповідного значення в отримане рівняння. В *алгоритмічному підході* передбачається алгоритм розрахунку невипадкової складової в будь-який заданий момент часу. До алгоритмічного підходу відносять ковзні середні, які дозволяють згладити як випадкові, так і періодичні коливання, виявити наявну тенденцію у розвитку процесу.

Алгоритм згладжування за простою ковзною середньою реалізується за такими етапами:

1) визначають довжину інтервалу згладжування l, що включає в себе l послідовних рівнів ряду ($l < n$). Чим ширший інтервал згладжування, тим більшою мірою поглинаються коливання, і тенденція розвитку носить більш

плавний характер. Чим сильніше коливання, тим ширший повинен бути інтервал згладжування;

2) розбивають весь період спостережень на частини й інтервал згладжування переходить рядом з кроком, що дорівнює 1;

3) розраховують середні арифметичні з рівнів ряду, що утворюють кожну частину;

4) замінюють фактичні значення ряду, що знаходяться в центрі кожної частини, на відповідні середні значення.

Рекомендують довжину інтервалу згладжування l вибирати непарним числом $l = 2p + 1$, оскільки в цьому випадку отримані значення ковзної середньої припадають на середній член інтервалу. Спостереження, які беруть для розрахунку середнього значення, є *активною частиною*. За непарним значенням $l = 2p + 1$ всі рівні активної частини можуть бути подані у вигляді:

$$y_{t-p}, y_{t-p+1}, ..., y_{t-1}, y_t, y_{t+1}, ..., y_{t+p-1}, y_{t+p}, \tag{9.3}$$

де y_t – центральний рівень активної частини;

$y_{t-p}, y_{t-p+1}, ..., y_{t-1}$ – послідовність з p рівнів активної частини, попередньої центральної;

$y_{t+1}, ..., y_{t+p-1}, y_{t+p}$ – послідовність з рівнів активної частини, що йде за центральною.

У цьому випадку ковзна середня визначається за формулою:

$$\overline{y_t} = \frac{\sum_{i=t-p}^{t+p} y_i}{2p+1} = \frac{y_{t-p} + y_{t-p+1} + ... + y_{t+p-1}, y_{t+p}}{2p+1}, \tag{9.4}$$

де y_i – фактичне значення *i*-го рівня; $\overline{y_t}$ – значення ковзної середньої в момент t; $2p+1$ – довжина інтервалу згладжування.

У ході використання простої ковзної середньої вирівнювання в кожній активній частині проводиться на прямій і апроксимація невипадкової складової здійснюється за допомогою лінійної функції $\overline{y_t} = a_0 + a_1 t$.

Метод згладжування призводить до усунення періодичних коливань у часовому ряді, якщо довжина інтервалу згладжування дорівнює або кратна періоду коливань. Рекомендується для усунення сезонних коливань використовувати

ковзні середні з довжиною інтервалу згладжування, що дорівнюють 4 або 12, проте тоді не буде виконуватися умова непарності. У цьому випадку ковзна середня розраховується за формулою:

$$\overline{y_t} = \frac{\frac{1}{2}y_{t-p} + y_{t-p+1} + \ldots + y_{t-1} + y_t + y_{t+1} + \ldots + y_{t+p-1} + \frac{1}{2}y_{t+p}}{2p} =$$

$$= \frac{\frac{1}{2}y_{t-p} + \sum_{i=t-p+1}^{t+p-1} y_i + \frac{1}{2}y_{t+p}}{2p}. \qquad (9.5)$$

Тому для згладжування сезонних коливань з квартальним або місячним часовими рядами використовують 4-членну і 12-членну ковзну середню:

$$\overline{y_t} = \frac{\frac{1}{2}y_{t-2} + y_{t-1} + y_t + y_{t+1} + \frac{1}{2}y_{t+2}}{4};$$

$$\overline{y_t} = \frac{\frac{1}{2}y_{t-6} + y_{t-51} + \ldots + y_t + \ldots + y_{t+5} + \frac{1}{2}y_{t+6}}{12}. \qquad (9.6)$$

Метод простої ковзної середньої рекомендується застосовувати, якщо графічне зображення часового ряду нагадує пряму. Якщо ж даний процес розвивається нелінійно, то проста ковзна середня призводить до істотних спотворень, і в цьому випадку рекомендують використовувати зважену ковзну середню:

$$\overline{y_t} = \frac{\sum_{i=t-p}^{t+p} y_i w_i}{\sum_{i=t-p}^{t+p} w_i}, \qquad (9.7)$$

де w_i – вагові коефіцієнти.

Вагові коефіцієнти мають властивості:

1) симетричні щодо центрального рівня;

2) сума ваги з урахуванням загального множника, винесеного за дужки, дорівнює одиниці;

3) є як позитивна, так і негативна вага, що дозволяє згладженій кривій зберігати різні згини кривої тренду.

9.2. Теоретичні відомості про функції *MATLAB*, які використовуються в даній лабораторній роботі

Econometrics Toolbox **– набір інструментів моделювання на основі принципів поведінки економічних систем.**

Econometrics Toolbox містить функції для моделювання економічних процесів. Надається набір моделей, які можна калібрувати за даними та використати для симуляцій та прогнозування.

Засоби аналізу часових рядів включають одномірні *ARMAX/GARCH* моделі, багатомірні *VARMAX* моделі, коінтеграцію, тести для вибору моделей: корінь із одиниці, тест стаціонарності. Також *ToolBox* надає методи Монте-Карло для розв'язання лінійних і нелінійних систем стохастичних диференціальних рівнянь.

Econometrics Toolbox підтримує ітеративний процес ідентифікації та тестування одно- і багатомірних фінансових й економічних часових рядів.

ToolBox організує повний цикл від розроблення до аналізу моделі:

аналіз даних і передоброблення;

ідентифікація моделі;

оцінювання параметрів;

симуляція;

прогнозування.

Засоби аналізу часових рядів в *Econometrics Toolbox* дозволяють моделювати більшість характеристик, що асоціюються з фінансовими й економічними рядами, включаючи тяжкі хвости, кластеризацію волатильності й ефекти кредитного плеча.

Моделі умовного середнього:

Autoregressive moving average (*ARMA*);

Autoregressive moving average з зовнішніми даними (*ARMAX*).

Моделі умовної волатильності:

узагальнена авторегресійна умовна гетероскедастичність (*GARCH*);

Glosten-Jogannathan-Runkle (GJR);

Exponential GARCH (*EGARCH*).

`SMOOTH` **Згладжування даних**

Синтаксис

```
Z = SMOOTH(Y)
Z = SMOOTH(Y,SPAN)
Z = SMOOTH(Y,SPAN,METHOD)
```

Опис

Z = SMOOTH(Y) згладжує дані *Y* використовуючи метод ковзного середнього значення на 5 пунктів.

Z = SMOOTH(Y,SPAN) згладжує дані *Y* з кроком *SPAN*.

Z = SMOOTH(Y,SPAN,METHOD) згладжує дані *Y* із заданим методом. Доступні методи:

'moving' – ковзне середнє значення (за замовчуванням);

'lowess' – ловес (лінійне припасування)

'loess' – лес (квадратне припасування)

'sgolay' – згладжування Савицького – Голея

'rlowess' – робастний ловесс (лінійне припасування)

'rloess' – робастний лес (квадратичне припасування)

Z = SMOOTH(Y,METHOD) згладжує дані *Y* із заданим методом, значення *SPAN* за замовчуванням дорівнює 5.

Z = SMOOTH(Y,SPAN,'sgolay',DEGREE) і *Z = SMOOTH(Y,'sgolay',DEGREE)* додатково дозволяє задавати ступінь полінома, що буде використовуватися в методі *Savitzky-Golay*. Ступінь за замовчуванням дорівнює 2. Ступінь повинна бути меншої, ніж проміжок.

Z = SMOOTH(X,Y,...) додатково визначає *X* координати. Якщо *X* не заданий, методи, які вимагають завдання координат, приймають *X = 1:N*, де *N* – довжина *Y*.

Примітки:

1. Коли *X* заданий і *X* є однорідно нерозподілений, метод за замовчуванням – *'lowess'*. Метод ковзного середнього застосовувати не рекомендується.

2. Для методів *'moving'* і *'sgolay'*, *SPAN* повинен бути непарним. Якщо *SPAN* визначений парним, він буде зменшений на 1.

3. Якщо *SPAN* більше довжини *Y*, його довжина зменшується до довжини *Y*.

4. У випадку (робастного) *lowess* і (робастного) *loess*, також можливо визначити *SPAN* як відсоток загальної кількості точок даних. Коли *SPAN* менше або дорівнює 1, його розглядають як відсоток.

9.3. Розв'язування типової задачі в середовищі *MATLAB*

Задача. Дано обсяги товарообігу фірми за три роки поквартально (табл. 9.1).

Таблиця 9.1

Вихідні дані задачі

Рік	**1994**				**1995**				**1996**			
Квартал	I	II	III	IV	I	II	III	IV	I	II	III	IV
Обсяг товарообігу, тис. шт.	16.8	16.2	14.7	15.8	15.8	15.4	16.3	15.8	17.9	18.5	21.2	19.3

Для початку роботи необхідно створити новий файл. Для цього з меню File вибрати опцію New, а потім M-File. У вікні, що з'явилося, редактора М-файлів ввести вихідні дані.

Фрагмент М-файла

```
% Аналіз часових рядів
clc
clear all
% Вхідні дані
t = [1 2 3 4 5 6 7 8 9 10 11 12]';
y = [16.8 16.2 14.7 15.8 15.8 15.4 16.3 15.8 17.9 18.5 21.2
19.3]';
```

Використовуючи функцію *smooth* здійснити згладжування часового ряду. У якості ковзної середньої *MATLAB* використовує середнє арифметичне за рівнями часового ряду. За допомогою функції *plot* зобразити вихідні дані (змінна *y*) і ряд ковзних середніх (змінна *ys*) (рис. 9.1):

Фрагмент М-файла

```
% Метод ковзних середніх
ys = smooth(y,3)
plot(t,y,'b-',t,ys,'r-')
grid on
hold on
title ('Згладжування ряду')
xlabel('t')
```

```
ylabel('y')
legend('Вхідні дані', 'Ряд ковзних середніх')
```

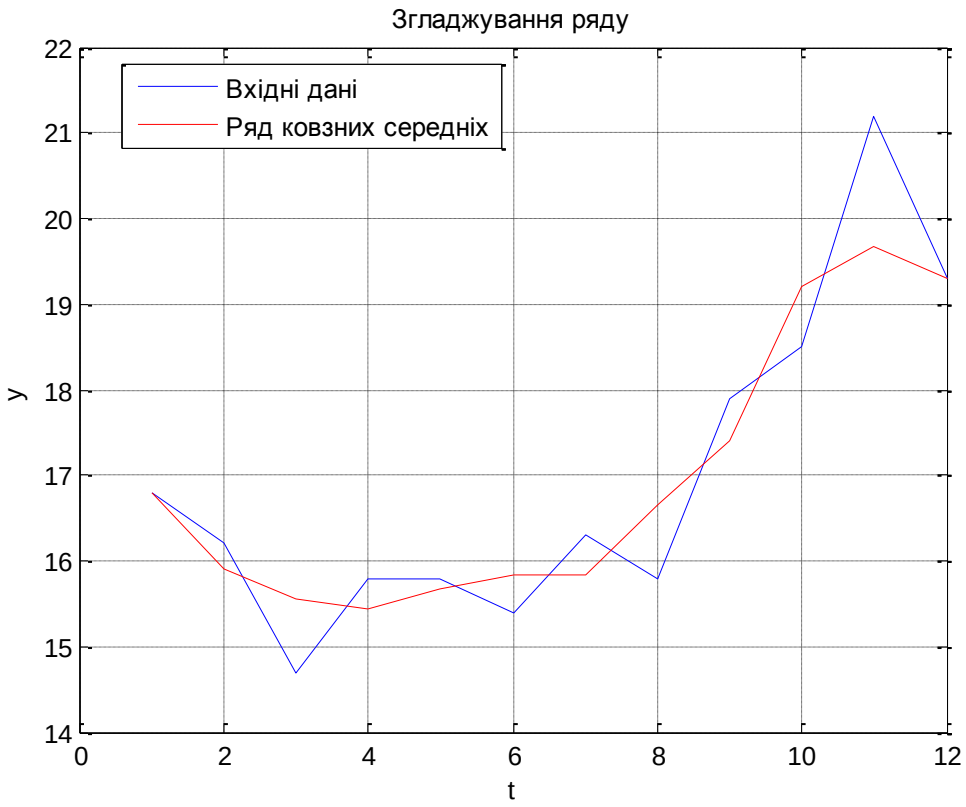


Рис. 9.1. **Графічне зображення вхідних даних та ряду ковзних середніх**

```
>> ys =

   16.8000
   15.9000
   15.5667
   15.4333
   15.6667
   15.8333
   15.8333
   16.6667
   17.4000
   19.2000
   19.6667
   19.3000
```

Отриманий ряд ковзних середніх становить часовий ряд з вилученою циклічною складовою.

Для того щоб визначити вид моделі часового ряду (адитивна або мультиплікативна), ввести нову змінну, котра буде містити циклічну складову (C):

Фрагмент М-файла

```
% Циклічна складова
C = y - ys;
[y ys C]
plot(t, C)
grid on
hold on
title ('Циклічна складова')
xlabel('t')
ylabel('C')
```

Оскільки циклічна складова має періодичний характер зміни свого значення, то слід вибрати адитивну модель часового ряду.

Необхідно виділити й усунути сезонні коливання з часового ряду (у даній задачі сезонність є річна; специфіка економічного розвитку також розглядається як річна). Для цього треба створити нову змінну Cp і записати в неї середні значення циклічної складової за кварталами (функція *mean* середовища *MATLAB*) .

Фрагмент М-файла

```
% Абсолютне відхилення в сезоні:
% Обчислюємо абсолютне відхилення по кожному року, усереднюючи
циклічну складову, тобто одержуємо оцінку сезонної складової
за відповідним роком
Cp = [mean(C(2:4)) mean(C(5:8)) mean(C(9:11))]
if sum(Cp) == 0
    fprintf('\nсума всіх сезонних компонент дорівнює нулю\n')
else
% Оскільки вимога до сезонних складових не виконується, то
розглядати будемо виправлене абсолютне відхилення за кожним роком
    fprintf('\nсума всіх сезонних компонент не дорівнює нулю\n')
    fprintf('\n виправлене абсолютне відхилення за кожним роком
\n')
    Cp_is = [Cp(1)-mean(Cp) Cp(2)-mean(Cp) Cp(3)-mean(Cp)]
end
```

```
>> Cp =

   -0.0667   -0.1750    0.4444
```

Сума всіх сезонних компонент не дорівнює нулю

Оскільки вимога до сезонних складових не виконується, то розглядати слід виправлене абсолютне відхилення за кожним роком:

```
>> Виправлене абсолютне відхилення за кожним роком

Cp_is =

   -0.1343   -0.2426    0.3769
```

Таким чином, змінна *Cp_is* містить оцінки сезонних складових з урахуванням вимог до них.

Слід доповнити аналіз часового ряду змінною *Y*, яка буде містити ряд з вилученим сезонним компонентом. Видаляти сезонну складову треба, віднімаючи з рівнів вихідного часового ряду відповідну оцінку сезонної складової.

Фрагмент М-файла

```
fprintf('\nРяд з видаленою сезонною компонентою (Y)\n')
Y = zeros(length(y),1);
for i = 1:4
    Y(i) = y(i) - Cp_is(1);
end
for i = 5:8
    Y(i) = y(i) - Cp_is(2);
end
for i = 9:12
    Y(i) = y(i) - Cp_is(3);
end
[y Y]
```

```
>> Ряд з видаленою сезонною компонентою (Y) у зіставленні
з вихідним рядом:

   16.8000   16.9343
   16.2000   16.3343
   14.7000   14.8343
   15.8000   15.9343
```

```
   15.8000   16.0426
   15.4000   15.6426
   16.3000   16.5426
   15.8000   16.0426
   17.9000   17.5231
   18.5000   18.1231
   21.2000   20.8231
   19.3000   18.9231
```

Необхідно визначити вид функції тренда. Для цього слід ввести три нових змінних: кінцеві різниці першого порядку (*del1*), кінцеві різниці другого порядку (*del2*) і темпи приросту (*temp*).

Фрагмент М-файла

```
fprintf('\nКінцеві різниці першого порядку\n')
del1 = zeros(length(y),1);
for i=2:length(y)
    del1(i) = Y(i) - Y(i-1);
end

fprintf('\nКінцеві різниці другого порядку\n')
del2 = zeros(length(del1),1);
for i=3:length(del1)
    del2(i) = del1(i) - del1(i-1);
end

fprintf('\nТемпи приросту\n')
temp = zeros(length(Y),1);
for i=2:length(Y)
    temp(i) = (Y(i) - Y(i-1))/Y(i-1);
end
[del1 del2 temp]
```

```
         0         0         0
   -0.6000         0   -0.0354
   -1.5000   -0.9000   -0.0918
    1.1000    2.6000    0.0742
    0.1083   -0.9917    0.0068
   -0.4000   -0.5083   -0.0249
    0.9000    1.3000    0.0575
   -0.5000   -1.4000   -0.0302
    1.4806    1.9806    0.0923
```

```
    0.6000   -0.8806    0.0342
    2.7000    2.1000    0.1490
   -1.9000   -4.6000   -0.0912
```

Порівнюючи отримані стовпці можна сказати, що тенденцію краще виражати показниковою регресією: $T = b_0 \cdot b_1^t$.

Використовуючи функцію *cftool*, оцінити параметри тренда й усунути останній з часового ряду (рис. 9.2):

```
>> cftool
```

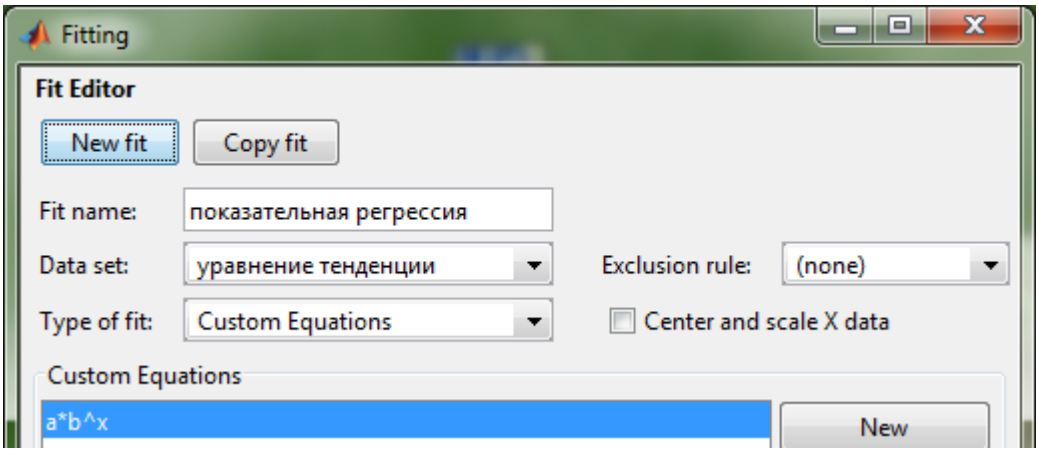


Рис. 9.2. **Побудова регресійного рівняння за допомогою функції *cftool***

```
General model:
     f(x) = a*b^x

Coefficients (with 95% confidence bounds):
       a =       14.85  (13.35, 16.34)
       b =        1.02  (1.007, 1.034)

Goodness of fit:
  SSE: 13.84
  R-square: 0.5386
  Adjusted R-square: 0.4925
  RMSE: 1.176
```

MATLAB автоматично обчислює параметри моделі разом із довірчими інтервалами для кожного з параметрів (95 % інтервал за замовчуванням),

коефіцієнт детермінації та скоригований коефіцієнт детермінації, а також іншу інформацію (опис вихідних параметрів функції *cftool* наведений у лабораторній роботі 3 даного посібника), що зберігається у двох структурах: *goodness*1 та *output*1.

Необхідно записати рівняння тенденції, а також отримати ряд залишків, видаляючи значення тренда з вихідного часового ряду:

Фрагмент М-файла

```
b = coeffvalues(fittedmodel1);
fprintf('\nрівняння тенденції\n')
fprintf('\n Yp = %f * %f^t\n',b)
Yp = b(1) * b(2).^t;
fprintf('\nРяд залишків\n')
r = Y - Yp
```

```
>> Рівняння тенденції

 Yp = 14.846372 * 1.020429^t

>> Ряд залишків

r =
    1.7846
    0.8751
   -0.9407
   -0.1630
   -0.3835
   -1.1191
   -0.5615
   -1.4110
   -0.2870
   -0.0508
    2.2779
   -0.0010
```

Тепер, отримавши ряд залишків часового ряду, можна провести аналіз якості побудованої моделі часового ряду, а щодо вихідного часового ряду можна сказати, що він став стаціонарним і становить тепер ряд залишків часового ряду.

Проаналізуємо якість побудованої моделі часового ряду.

Фрагмент М-файла

```
% Аналіз моделі:
fprintf('Коефіцієнт детермінації дорівнює')
```

```
R2 = goodness1.rsquare
fprintf('Скорегований коефіцієнт детермінації дорівнює')
R2_kop = goodness1.adjrsquare
fprintf('Критерій Фішера')
n = goodness1.dfe;
m = 1;
F = (R2/(1-R2))*(n - m - 1)% F-статистика
Ft = finv(0.95,m,n-m-1)
if F > Ft
    fprintf('Модель значуща в цілому\n')
else
    fprintf('\nМодель не є значущою')
end
```

```
>> Коефіцієнт детермінації дорівнює

R2 =    0.5386
>> Скорегований коефіцієнт детермінації дорівнює
R2_kop =  0.4925

>> Критерій Фішера

F =

    9.3400

Ft =

    5.3177

Модель значуща в цілому
```

Отже, можна сказати, що включені до моделі фактори пояснюють 53,86 % загальної варіації рівнів часового ряду. Оскільки $F > F_t$, то рівняння статистично значуще. Можна вважати, що побудована модель часового ряду якісна, тобто вона адекватно описує вихідні дані.

Використовуючи вбудовану функцію *autocorr*, створити автокореляційну функцію та побудувати корелограму:

Фрагмент М-файла

```
autocorr(y)
[ACF, Lags, Bounds] = autocorr(y)
```

```
>> ACF =

    1.0000
    0.6675
    0.3437
    0.0400
   -0.1203
   -0.1941
   -0.2923
   -0.2970
   -0.3481
   -0.2246
   -0.0646
   -0.0103
>> Lags =

     0
     1
     2
     3
     4
     5
     6
     7
     8
     9
    10
    11

>> Bounds =

    0.5774
   -0.5774
```

Корелограма має вигляд (рис. 9.3).

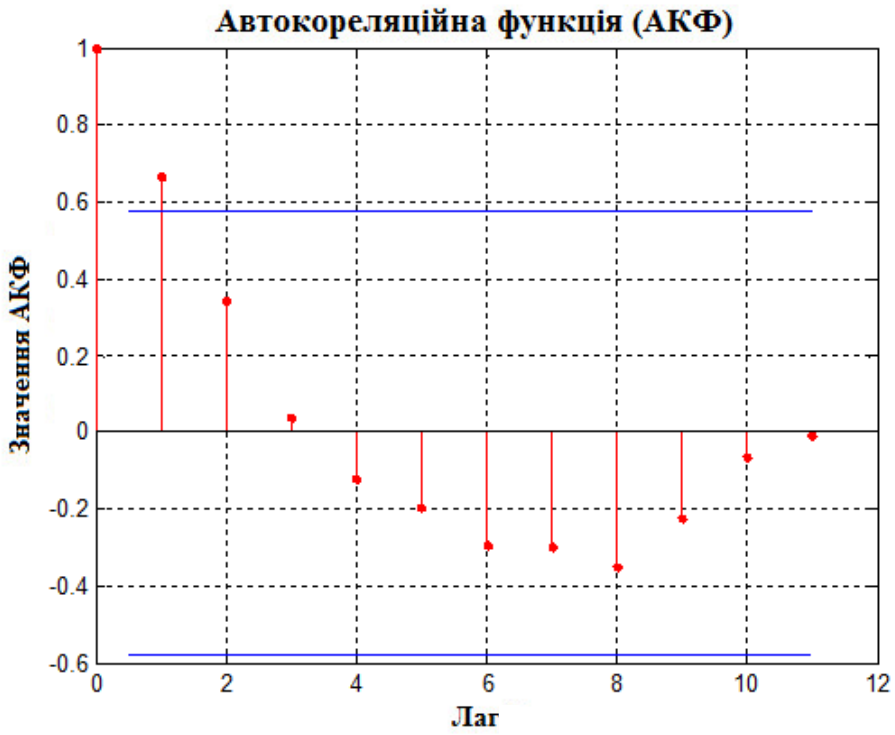


Рис. 9.3. **Корелограма моделі часового ряду**

Точковий прогноз обсягу товарообігу на перший квартал 1997 року здійснити за побудованою моделлю, підставивши номер кварталу в рівняння тренда та визначивши значення залежної змінної `Yp`:

Фрагмент М-файла

```
%Прогноз
t = 13;
Yp = b(1) * b(2).^t
```

```
Yp =

   19.3107
```

Таким чином, обсяг товарообігу на перший квартал 1997 року складе 19,31 тис. шт.

Дати графічну інтерпретацію побудованої моделі часового ряду, зобразивши на одній площині оцінку тренда, вихідний часовий ряд і ряд, очищений від впливу циклічності та сезонності (рис. 9.4).

Фрагмент М-файла

```
t = [1 2 3 4 5 6 7 8 9 10 11 12]';
Yp = b(1) * b(2).^t;
plot(t,y,'r-',t,Y,'g-',t,Yp)
grid on
hold on
title ('Модель часового ряду "Обсяг товарообігу"')
xlabel('t')
ylabel('Обсяг товарообігу')
legend('Вихідні дані','Ряд, очищений від впливу циклічності та сезонності','Оцінка тренда')
```

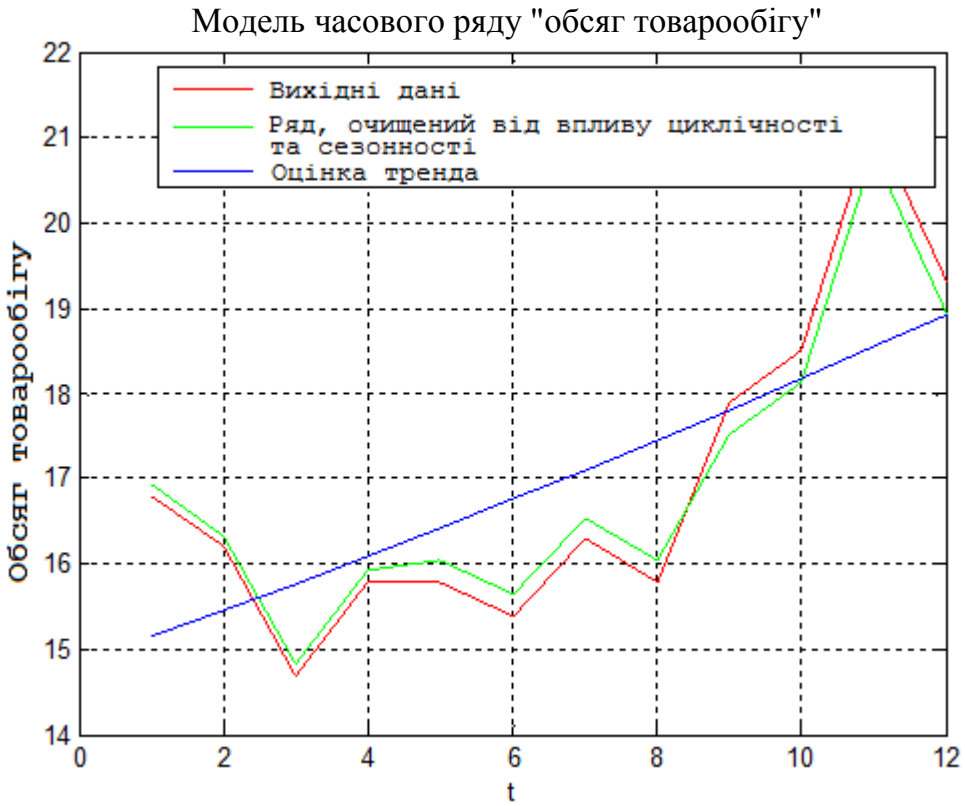


Рис. 9.4. **Моделі часового ряду "Обсяг товарообігу"**

Запитання для самоперевірки

1. Які основні причини використання систем одночасних рівнянь?
2. Наведіть основні типи моделей застосування систем одночасних рівнянь.
3. Які види змінних розрізняють в системах одночасних рівнянь? Дайте їм визначення.
4. У чому полягає відмінність між структурними рівняннями системи та рівняннями у приведеній формі?

5. Чому звичайний МНК практично не використовується для оцінювання систем одночасних рівнянь?

6. Які найбільш поширені методи оцінювання коефіцієнтів структурної моделі?

7. У чому полягає суть непрямого методу найменших квадратів?

8. Назвіть причини неідентифікованості та надідентифікованості систем одночасних рівнянь.

9. Наведіть необхідні та достатні умови ідентифікованості систем.

10. У чому полягає сутність двокрокового методу найменших квадратів?

Завдання для лабораторної роботи

Задача. Вихідні дані валового регіонального продукту (до 1998 р. у млрд у. о., з 1998 р. – у млн у. о.) наведені в табл. 9.2 (p_1 – число букв у повному імені, p_2 – число букв у прізвищі).

Таблиця 9.2

Вихідні дані задачі

Роки	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
ВРП, млн у. о.	6 135 $+ p_1$	15074 $+ p_1$	23909 $+ p_1$	30866 $+ p_1$	29879 $- p_2$	41786 $- p_2$	67835 $- p_2$	81961 $- p_2$	101848 $- p_2$	122550 $- p_2$

Виходячи з даних табл. 9.2 провести наступні обчислення.

1. За вихідним даними зробити згладжування ряду (усунути циклічні коливання з часового ряду).

2. Визначити вид моделі часового ряду (адитивна або мультиплікативна).

3. Виділити й всунути сезонні коливання з часового ряду.

4. Визначити вид функції тренда.

5. Оцінити параметри тренда та усунути його з часового ряду.

6. Використовуючи вбудовану функцію `autocorr` розрахувати коефіцієнти кореляції та побудувати корелограму.

7. Здійснити прогноз обсягу товарообігу на перший квартал 1997 року.

8. Дати графічну інтерпретацію побудованої моделі часового ряду.

Кожна лабораторна робота повинна бути окремим робочим модулем, написаним у М-файлі.

Використана література

1. Айвазян С. А. Прикладная статистика и основы эконометрики : учебник для вузов / С. А. Айвазян, В. С. Мхитарян. – М. : ЮНИТИ-ДАНА, 1998. – 1022 с.

2. Доугерті К. Введение в эконометрику / К. Доугерти; пер. с англ. – М. : ИНФРА-М, 1999. – 402 с.

3. Єгоршин О. О. Лабораторний практикум з навчальної дисципліни "Економіко-математичні методи та моделі: економетрика" : навч.-практ. посіб. / О. О. Єгоршин, Л. М. Малярець. – Х. : Вид. ХНЕУ, 2011. – 148 с.

4. Магнус Я. Р. Эконометрика : начальный курс / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – М. : Дело, 1997. – 248 с.

5. Малярец Л. М. Экономико-математические методы и модели : учеб. пособ. для иностранных студентов / Л. М. Малярец. – Х. : Изд. ХНЭУ, 2013. – 288 с.

Зміст

НАВЧАЛЬНЕ ВИДАННЯ

Малярець Людмила Михайлівна
Ковальова Катерина Олександрівна

Лабораторний практикум з навчальної дисципліни "ЕКОНОМЕТРИКА" В СЕРЕДОВИЩІ MATLAB

Навчальний посібник

Відповідальний за випуск *Малярець Л. М.*

Відповідальний редактор *Оленич М. М.*

Редактор *Ганцевич Н. І.*

Коректор *Лященко О. Г.*

План 2015 р. Поз. № 21-П.
Підп. до друку 24.12.2015 р. Формат 60 х 90 1/16. Папір офсетний. Друк цифровий.
Ум. друк. арк. 12,0. Обл.-вид. арк. 15,0. Тираж 400 пр. Зам. № 269.

Видавець і виготівник – ХНЕУ ім. С. Кузнеця, 61166, м. Харків, просп. Леніна, 9-А

Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру
ДК № 4853 від 20.02.2015 р.