МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ І ПРИКЛАДНИХ ТЕХНОЛОГІЙ

Л. О. Волонтир, Н. А. Потапова, Ю. С. Хмелівський

СТАТИСТИЧНЕ НАВЧАННЯ. ЧАСТИНА 1

Методичні вказівки
для виконання лабораторних робіт
для здобувачів ОС «Бакалавр»
спеціальності 122 Комп’ютерні науки

Вінниця
2024

УДК 004.9(076.5)
В 682

Затверджено на засіданні вченої ради факультету інформаційних і прикладних технологій ДонНУ імені Василя Стуса (протокол № 1 від 28 серпня 2024 р.)

Автори:

Волонтир Л. О., канд. техн. наук, доцент кафедри інформаційних технологій ДонНУ імені Василя Стуса;

Потапова Н. А., канд. екон. наук, доцент кафедри інформаційних технологій ДонНУ імені Василя Стуса;

Хмелівський Ю. С., асистент кафедри інформаційних технологій ДонНУ імені Василя Стуса.

Рецензенти:

Денисюк В. О., канд. техн. наук, доцент, доцент кафедри комп'ютерних наук ВНТУ;

Бабаков Р. М., д-р техн. наук, доцент, професор кафедри інформаційних технологій ДонНУ імені Василя Стуса.

Волонтир Л. О., Потапова Н. А., Хмелівський Ю. С.

В 682 Статистичне навчання. Частина 1: методичні вказівки для виконання лабораторних робіт для здобувачів ОС «Бакалавр» спеціальності 122 Комп'ютерні науки. Вінниця, ДонНУ імені Василя Стуса, 2024. 48 с.

Методичні вказівки є навчально-методичним документом під час виконання лабораторного практикуму з дисципліни «Статистичне навчання» для здобувачів вищої освіти ОС «Бакалавр» спеціальності 122 Комп'ютерні науки факультету інформаційних і прикладних технологій ДонНУ імені Василя Стуса.

УДК 004.9(076.5)

ЗМІСТ

ВСТУП

До статистичного навчання належить набір інструментів, призначених для моделювання та розуміння складно організованих даних. Це нещодавно розроблена галузь статистики, яка розвинулася паралельно з досягненнями в комп'ютерних науках і особливо машинному навчанні. Ця галузь охоплює багато методів, зокрема ласо і розріджену регресію, класифікаційні та регресивні дерева, бустінг і метод опорних векторів.

Дисципліна «Статистичне навчання» продовжує цикл дисциплін, пов'язаних із вивченням математичних і статистичних методів моделювання й прогнозування розвитку складних систем та їх практичного використання у професіональній діяльності. Знання курсу дає змогу розв'язувати задачі прогнозування і статистичного аналізу складних систем із застосуванням сучасних програмних засобів.

МЕТА ТА ЗАВДАННЯ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Дисципліна «Статистичне навчання» має на меті ознайомлення студентів із загальними принципами статистичного навчання, видами, моделями, методами статистичного навчання та прикладними задачами, які розв'язуються на основі цих методів і моделей.

Завданням навчальної дисципліни є набуття таких результатів навчання:

- знати методи, моделі та галузі практичного застосування: навчання з вчителем і без вчителя; відмінності між проблемами регресії і класифікації; вимірювання якості моделі; проста лінійна регресія; оцінювання точності моделі; множинна лінійна регресія; якісні предиктори; розширення лінійної моделі;
- загальне уявлення про класифікацію; логістична регресія; логістична модель; логістична регресія для залежних змінних;
- дискримінантний аналіз; використання теореми Баєса для класифікації; лінійний дискримінант аналіз для $p = 1$; лінійний дискримінантний аналіз для $p > 1$; квадратичний дискримінантний аналіз;
- перехресна перевірка; метод перевірочної вибірки; перехресна перевірка по окремим спостереженням; k-кратна перехресна перевірка; перехресна перевірка під час розв'язання задач класифікації; бутстреп;
- відбір підмножини змінних; відбір оптимальної підмножини; покроковий відбір; вибір оптимальної моделі; методи стиснення; гребенева регресія; ласо; вибір гіперпараметра;
- методи зниження розмірності; регресія на головні компоненти; метод частинних найменших квадратів; особливості роботи з даними великої розмірності; регресія для даних великої розмірності; інтерпретація результатів у задачах великої розмірності; нелінійна регресія; локальна регресія; узагальнені адитивні моделі;
- поліноміальної регресія; ступінчасті функції; базисні функції; регресивні сплайни; кусково-поліноміальна регресія; обмеження і сплайни; подання сплайнів за допомогою базисних функцій; вибір числа і розташування вузлів зчленування; порівняння з поліноміальною регресією.

Внаслідок опанування дисципліни студенти мають вміти: будувати і досліджувати моделі регресії таких типів: лінійна, нелінійна, з кількісними і якісними предикторами, логістична; застосовувати методи класифікації; виконувати лінійний та нелінійний дискримінантний аналіз; застосовувати методи перехресної перевірки; метод перевірочної вибірки; виконувати відбір підмножини змінних; застосовувати методи зниження розмірності; використовувати такі види нелінійних моделей: поліноміальна регресія, ступінчасті функції, базисні функції, регресивні сплайни, кусково-поліноміальна регресія; використовувати дерева рішень для розв'язання задач класифікації.

Навчальна дисципліна формує міждисциплінарні взаємозв'язки з іншими дисциплінами, як-от: «Програмування», «Бази даних та інформаційні системи», «Теорія ймовірностей і математична статистика», «Аналіз даних». Опанування навчальної дисципліни передбачає формування та розвиток у здобувачів інтегральної, загальних та спеціальних компетентностей і результатів навчання відповідно до освітньої програми «Комп'ютерні науки» спеціальності 122 Комп'ютерні науки ОС «Бакалавр».

ЛАБОРАТОРНА РОБОТА № 1

Тема: виконання операцій з матрицями за допомогою функцій MS Excel. Обробка спостережень статистичних ознак. Знаходження значень статистичних критеріїв.

Мета: навчитись використовувати матричні функції та функції обробки масивів MS Excel для обробки спостережень статистичних ознак та знаходження значень статистичних критеріїв.

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

В Excel є три функції, призначені для роботи з матрицями, і всі вони входять до категорії Математичні:

- MMULT – обчислює добуток матриць;
- MINVERSE – обчислює матрицю, обернену до заданої;
- MDETERM – обчислює визначник матриці.

Аргументами всіх цих функцій є діапазони, що містять елементи матриць, по одному в кожній клітинці. Результатом виконання перших двох функцій є не окреме значення, а діапазон значень. Тому вводити їх потрібно так само, як і функцію FREQUENCY (ЧАСТОТА): треба виділити весь діапазон, де міститимуться результати, ввести формулу функції та натиснути клавіші Ctrl + Shift + Enter.

Операції множення матриці на число та додавання матриць у Microsoft Excel треба виконувати не за допомогою функцій, а з використанням формул. Якщо матриця множиться на число, то посилання на клітинку, де це число розміщене, має бути абсолютним, оскільки всі елементи матриці множитимуться на значення в тій самій клітинці. Під час додавання матриць варто використовувати відносні посилання.

Статистичні функції в Microsoft Excel

Статистична обробка даних – це збір, упорядкування, узагальнення та аналіз інформації з можливістю визначення тенденції і прогнозу досліджуваного явища. В Excel є величезна кількість інструментів, які допомагають проводити дослідження в цій галузі. Останні версії цієї програми у плані можливостей практично нічим не поступаються спеціалізованим програмам у галузі статистики. Головними інструментами для виконання розрахунків і аналізу є функції. Вивчимо загальні особливості роботи з ними, а також докладніше зупинимося на окремих найбільш корисних інструментах.

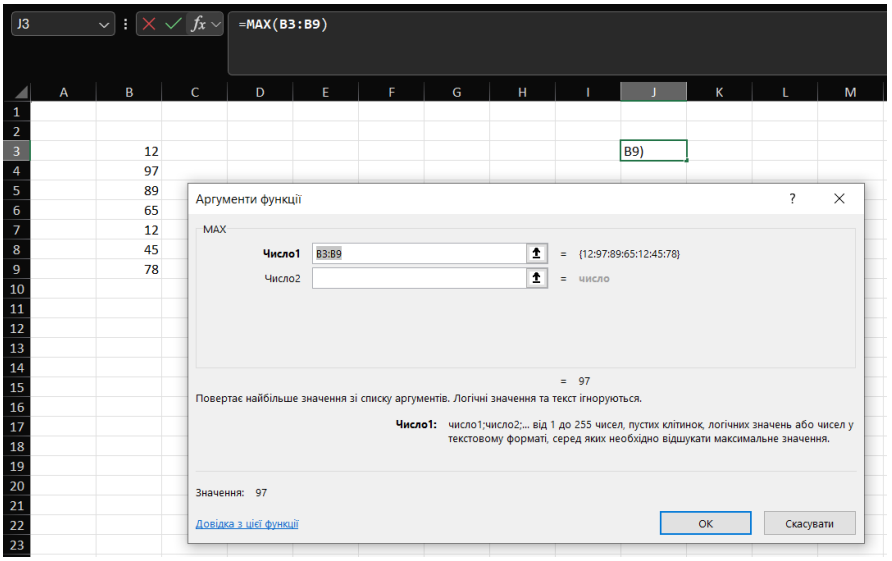
Як і будь-які інші функції в Excel, статистичні функції оперують аргументами, які можуть мати вигляд постійних чисел, посилань на осередки або масиви.

Вирази можна вводити вручну в певну комірку або в рядок формул, якщо добре знати синтаксис конкретного з них. Але набагато зручніше скористатися спеціальним вікном аргументів, яке містить підказки та вже готові поля для введення даних. Перейти у вікно аргументу статистичних виразів можна через *«Майстер функцій»* або за допомогою кнопок *«Бібліотеки функцій»* на стрічці.

MAX

Оператор MAX призначений для визначення максимального числа з вибірки. Він повинен мати такий вигляд:

```
=MAX(число1;число2;…)
```

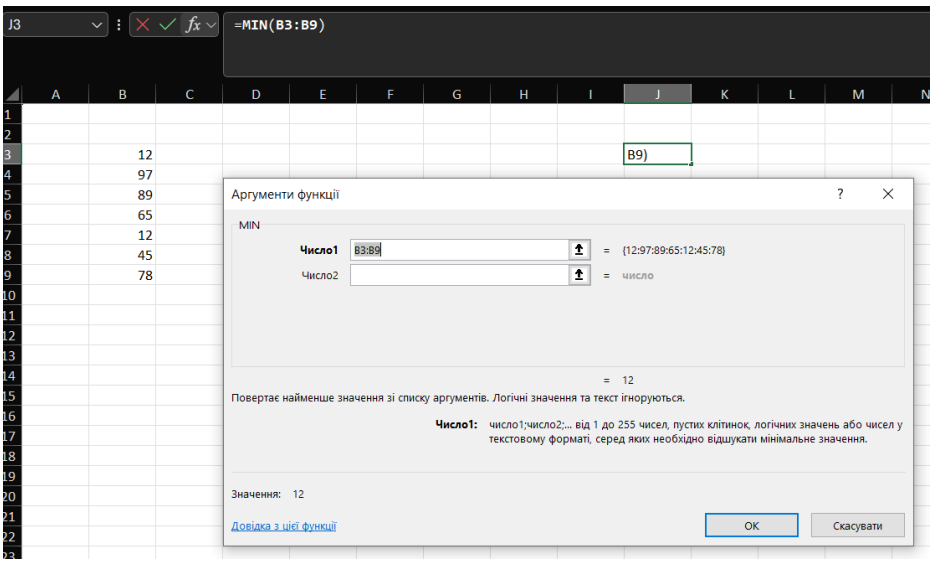


У поля аргументів потрібно ввести діапазони осередків, в яких знаходиться числовий ряд. Найбільше число з нього ця формула виводить у ту клітинку, в якій знаходиться сама.

MIN

За назвою функції MIN зрозуміло, що її завдання прямо протилежні попередній формулі – вона шукає з безлічі чисел найменше і виводить його в задану клітинку. Має такий синтаксис:

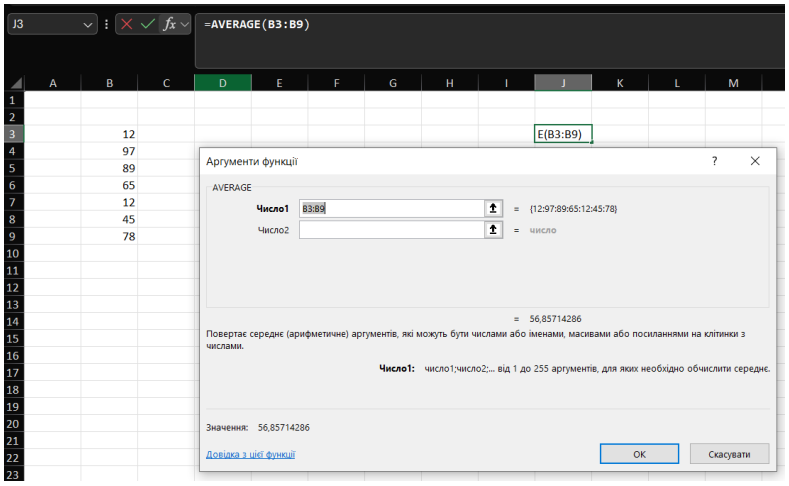
```
=MIN(число1;число2;…).
```



AVERAGE

Функція AVERAGE шукає середнє арифметичне значення. Результат цього розрахунку виводиться в окрему клітинку, в якій і міститься формула. Шаблон у неї такий:

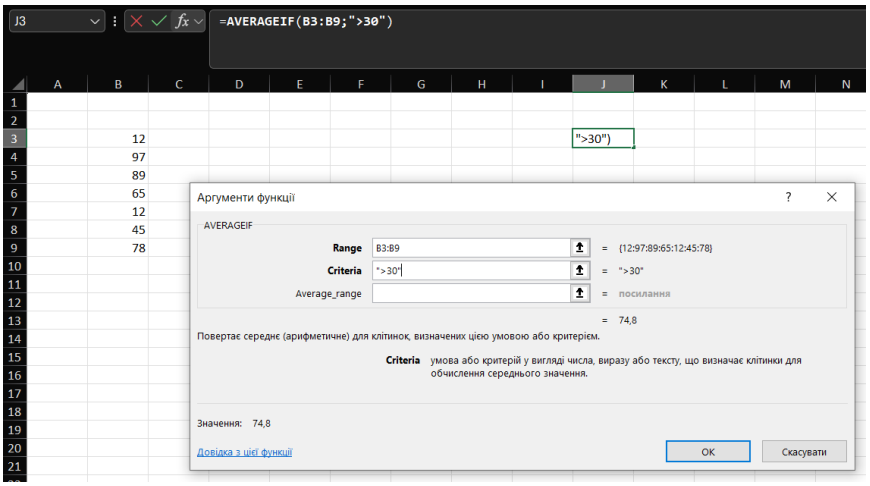
```
= AVERAGE(число1;число2;…).
```



AVERAGEIF

Функція AVERAGEIF має ті ж завдання, що і попередня, але в ній існує можливість задати додаткову умову. Наприклад, більше, менше, не дорівнює певному числу. Воно задається в окремому полі для аргументу. До того ж у якості необов’язкового аргументу може бути доданий діапазон усереднення. Синтаксис такий:

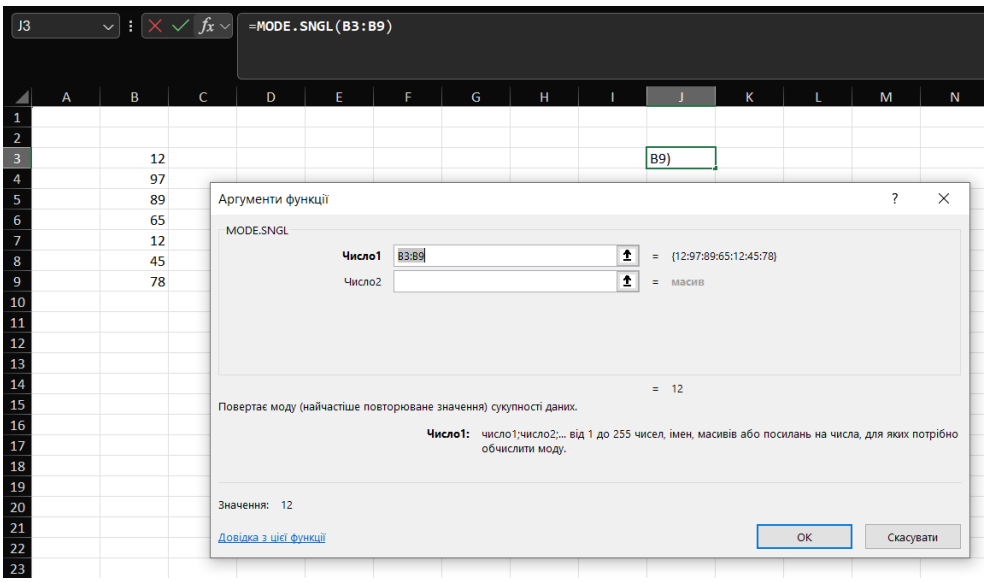
```
= AVERAGEIF(число1;число2;…;умова;[діапазон_усереднення])
```



MODE.SNGL

Формула MODE.SNGL виводить в осередок те число з набору, яке зустрічається найчастіше. У старих версіях Excel існувала функція MODE, але в більш пізніх вона була розбита на дві: MODE.SNGL (для окремих чисел) і МОДА.MULT (для масивів). Втім старий варіант теж залишився в окремій групі, в якій зібрані елементи з попередніх версій програми для забезпечення сумісності документів.

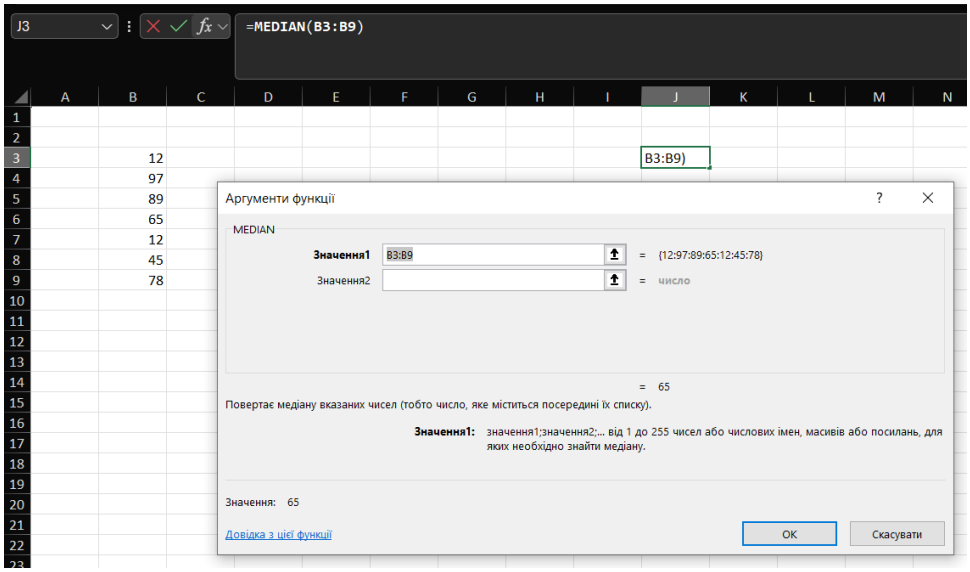
```
= MODE.SNGL (число1;число2;…);
= MODE.MULT (число1;число2;…).
```



MEDIAN

Оператор MEDIAN визначає середнє значення в діапазоні чисел, тобто встановлює не середнє арифметичне, а просто середню величину між найбільшим і найменшим числом області значень. Синтаксис має такий вигляд:

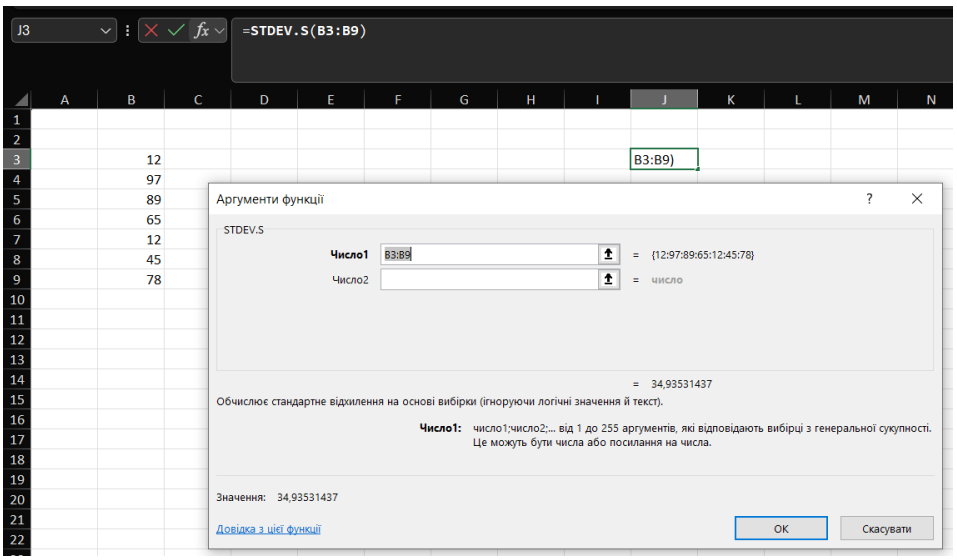
```
= MEDIAN (число1;число2;…).
```



STDEV.S

Формула STDEV так само, як і MODE є пережитком старих версій програми. Зараз використовуються сучасні її підвиди – STDEV.S і STDEV.P. Перша з них призначена для обчислення стандартного відхилення вибірки, а друга – генеральної сукупності. Ці функції використовуються також для розрахунку середнього квадратичного відхилення. Їх синтаксис такий:

```
= STDEV.S (число1;число2;…);
= STDEV.P (число1;число2;…).
```

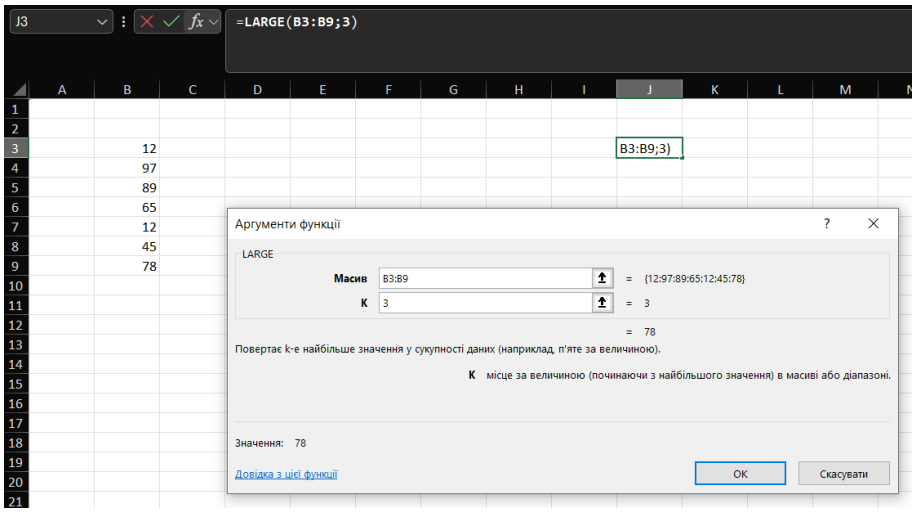


LARGE

Цей оператор показує у вибраній комірці вказане в порядку зменшення число зі сукупності. Тобто якщо ми маємо сукупність 12, 97, 89, 65, а аргументом позиції зазначимо 3, то функція в осередок поверне третє за величиною число. В цьому випадку це 65. Синтаксис оператора такий:

```
= LARGE (масив;k).
```

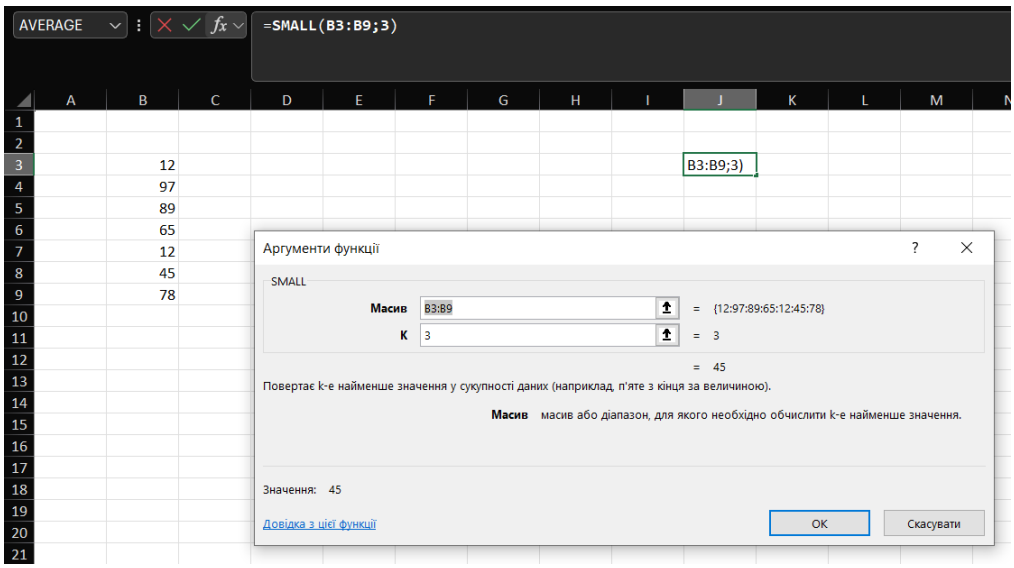
у цьому випадку k – порядковий номер величини.



SMALL

Ця функція є дзеркальним відображенням попереднього оператора. У ній також другим аргументом є порядковий номер числа. Тільки в цьому випадку порядок вважається від меншого. Синтаксис такий:

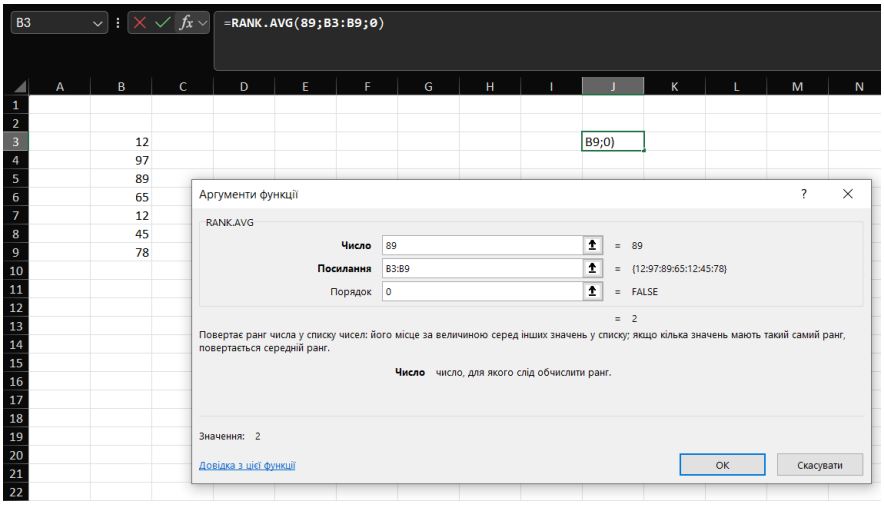
```
= SMALL (масив;k).
```



RANK.AVG

Ця функція має дію, зворотну попереднім. У зазначеному осередкові вона видає порядковий номер конкретного числа у вибірці за умовою, яка зазначена в окремому аргументі. Це може бути порядок за зростанням або за спаданням. Останній встановлений за замовчуванням, якщо поле *«Порядок»* залишити порожнім або вписати туди цифру 0. Синтаксис цього виразу має такий вигляд:

```
= RANK.AVG (число;масив;порядок).
```



Вище були описані тільки найпопулярніші і затребувані статистичні функції в Excel. Насправді їх у рази більше, проте основний принцип дій у них схожий – обробка масиву даних і повернення в зазначений осередок результату обчислювальних дій.

ЗАВДАННЯ ДО ЛАБОРАТОРНОЇ РОБОТИ

Завдання 1. Виконання операцій з матрицями за допомогою функцій MS Excel.

Для даної матриці $A = \begin{pmatrix} m+1 & m+2 & m+3 & m+4 \\ m+4 & m+5 & m+6 & m+7 \\ m+7 & m+8 & m+9 & m+9 \end{pmatrix}$,

де m – номер варіанта (відповідає порядковому номеру студента у списку групи). За допомогою вбудованих функцій програми MS Excel обчислити:

1) A^T;

2) $\mathrm{B} = AA^T$;

3) $\det B$;

4) B^{-1};

5) $m \cdot B^{-1}$.

Завдання 2. Обробка спостережень статистичних ознак.

За даними десяти спостережень статистичних ознак X і Y (табл. 1.1) за допомогою вбудованих функцій програми MS Excel:

- заповніть табл. 1.2 і визначте числові характеристики статистичних ознак X і Y;
- побудуйте кореляційне поле статистичних ознак X і Y.

Таблиця 1.1 – Спостереження статистичних ознак X і Y

X	$m+1$	$m+2$	$m+3$	$m+4$	$m+5$	$m+6$	$m+7$	$m+8$	$m+9$	$m+10$
Y	$0{,}2m$	$0{,}3m$	$0{,}3m$	$0{,}3m$	$0{,}4m$	$0{,}4m$	$0{,}4m$	$0{,}5m$	$0{,}5m$	$0{,}6m$

Завдання 3. Знаходження значень статистичних критеріїв.

Для заданих рівнів значимості $\alpha^1 = 0{,}05$ і $\alpha^1 = 0{,}01$ знайдіть:

- критичні (табличні) значення $F_{\text{табл.}}(\alpha, k_1, k_2)$ критерію Фішера (F-критерію) за умови ступенів свободи $k_1 = m+1$ і $k_2 = 50 - m$;
- критичні (табличні) значення $t_{\text{табл.}}(\alpha, k)$ критерію Стьюдента (t-критерію) за умови числа ступенів свободи $k = 50 - m$;
- критичні (табличні) значення ${\chi^2}_{\text{табл.}}(\alpha, k)$ χ-критерію за умови числа ступенів свободи $k = m$.

Таблиця 1.2 – Перетворення даних спостережень статистичних ознак X і Y

№ з/п	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{X^2}$	$\boldsymbol{Y^2}$	$\boldsymbol{XY}$	$\boldsymbol{X - X}$
1						
2						
…	…	…	…	…	…	…
n						
сума	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_i y_i$	–
середнє	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$	…	$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$	$\overline{Y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$	$\overline{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$	–
дисперсія	DX	DY	–	–	–	–
стандартне відхилення	σ X	σY	–	–	–	–
коваріація	cov(X, Y)		–	–	–	–
кореляція	r_{XY}		–	–	–	

КОНТРОЛЬНІ ПИТАННЯ

1. Правила виконання операцій з матрицями.
2. Порядок виконання операцій з матрицями в табличному редакторі MS Excel.
3. Назви функцій MS Excel виконання операцій з матрицями.
4. Вибіркові характеристики статистичних ознак. Розрахункові формули.
5. Функції MS Excel (категорія «Статистичні») для розрахунку вибіркових характеристик статистичних ознак.
6. Поняття табличного значення статистичного критерію (F-критерій, t-критерій, x^2-критерій) із заданим рівнем значущості і ступенями свободи.
7. Функції програми MS Excel, використовувані для знаходження значень F-критерію, t-критерію, x^2-критерію.

ЛАБОРАТОРНА РОБОТА № 2

Тема: вступ до пакету R. Основні команди. Графіки. Індексування та завантаження даних. Додаткові графічні та кількісні зведення.

Мета: ознайомлення з простими командами R.

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

У лабораторні роботі представлено деякі прості команди R. Найкращий спосіб вивчення нової мови полягає в експериментуванні з його командами. R можна завантажити з сайта: http://cran.r-project.org/

Основні команди

Для виконання тих чи інших операцій R використовує функції. Для запуску функції з ім'ям funcame ми набираємо funcname(input1, input2), де вхідні параметри, або аргументи input1 та input2, повідомляють R, як саме треба виконати цю функцію. Наприклад, для створення вектора з декількома числами ми використовуємо функцію конкатенації *c*(). Наступна команда говорить R об'єднати числа 1, 3, 2 та 5 і зберегти їх у вигляді вектору з ім'ям x. Коли ми наберемо x, то у відповідь отримаємо цей вектор:

```
>x <- c (1, 3, 2, 5)
>x
[1] 1 3 2 5
```

Зверніть увагу на те, що > не є частиною команди – R виводить цей знак просто, щоб показати свою готовність до виконання наступної команди. Ми можемо зберігати об'єкти не тільки за допомогою <-, але також і =:

```
>x = c (1, 6, 2)
>x
[1] 1 6 2
> y = c (1, 4, 3)
```

Багаторазове натискання клавіші зі стрілкою вгору призведе до показу попередніх команд, які можна відредагувати. Це корисно, оскільки необхідність повторення схожих команд виникає часто. До того ж введення команди **?funcname** завжди відкриває нове вікно з довідковим файлом, що містить додаткову інформацію по функції **funcname.**

Ми можемо попросити R виконати складання двох чисел. У цьому випадку програма спочатку добавить перше число з x до першого числа з y тощо. Однак x і y повинні бути однакової довжини. Ми можемо перевірити їх довжину за допомогою функції length():

```
>length (x)
[1] 3
>length (y)
[1] 3
>x + y
[1] 2 10 5
```

Функція **ls()** дає змогу переглянути список всіх об'єктів, як-от дані і функції, які ми зберегли до цього моменту. Функцію **rm()** можна використовувати для видалення будь-якого небажаного об'єкта:

```
>ls ()
[1] "x" "y"
```

```
>rm (x, y)
>ls () character (0)
```

Також є можливість видалити всі об’єкти за один раз:

```
>rm (list = ls ())
```

Функцію **matrix()** можна використовувати для створення матриці з числами. Перед застосуванням функції **matrix()** ми можемо дізнатися про неї більше:

```
>? matrix
```

Довідковий файл повідомляє, що функція **matrix()** приймає кілька вхідних параметрів, але поки ми зосередимося на перших трьох: дані (елементи матриці), кількість рядків і кількість стовпців. Спочатку створимо просту матрицю:

```
>x = matrix ( data=c(1, 2, 3, 4), nrow = 2, ncol = 2)
>x
[, 1] [, 2]
[1,] 1 3
[2,] 2 4
```

Зверніть увагу, що ми могли б із таким самим успіхом не набирати data =, nrow = та ncol = в наведеній вище команді matrix(), а просто ввести:

```
X = matrix (c (1, 2, 3, 4), 2, 2,)
```

І це мало б такий самий ефект. Однак іноді буває корисно вказати імена аргументів, інакше R буде припускати, що аргументи функції подаються на неї в тому ж порядку, який наведено в довідковому файлі цієї функції. Як показано в цьому прикладі, за замовчуванням R створює матриці шляхом послідовного заповнення стовпців. Як альтернативу можна використовувати опцію byrow = TRUE для заповнення матриці по рядках:

```
>matrix (c(1, 2, 3, 4), 2, 2, 2, byrow = TRUE)
[,1] [, 2]
[1, ] 1 2
[2, ] 3 4
```

Зауважте, що в наведеній вище команді ми не присвоїли матриці ніякого імені на кшталт x. У цьому випадку матриця виводиться на екран, але не зберігається для майбутніх обчислень. Функція **sqrt()** повертає квадратний корінь кожного елементу вектору або матриці. Команда *x^2* зводить кожен елемент *x* у квадрат; можливе використання будь-якого ступеня, зокрема дроби і негативні ступені:

```
>sqrt (x) [, 1] [, 2]
[1,] 1.00 1.73
[2,] 1.41 2.00
>x^2
[, 1] [, 2]
[1,] 1 9
```

```
[2,] 4 16
```

Функція rnorm() генерує вектор випадкових нормально розподілених значень, водночас аргумент n задає розмір вибірки. Щоразу під час виклику цієї функції ми будемо отримувати інший результат. Тут ми створюємо два корелювання набору чисел x і y та застосовуємо функцію cor() для розрахунку кореляції між ними:

```
>x= rnorm (50)
>y= x + rnorm (50, mean = 50, sd = .1)
>cor (x, y)
[1] 0.995
```

За замовчуванням rnorm() створює випадкові змінні, що представляють стандартний нормальний розподіл із середнім значенням 0 і стандартним відхиленням 1. Однак, як показано вище, середнє значення і стандартне відхилення можна змінити за допомогою аргументів mean та sd. Іноді ми хочемо, щоб наш код точно відтворював один і той самий набір випадкових чисел; для цього ми можемо використовувати функцію sed.seed(). Функція set.seed() приймає як аргумент довільне ціле число.

```
>set.seed (1303)
>rnorm (50)
[1] -1.1440 1.3421 2.1854 0.5364 0.0632 0.5022 -0.0004
. . .
```

Ми використовуємо set.seed() у всіх лабораторних роботах щоразу, коли виконуємо обчислення, що містять випадкові величини. Здебільшого це має допомогти користувачеві відтворити наші результати. Однак треба зазначити, що з появою нових версій R можуть виникнути невеликі невідповідності між книгою і тим, що видає R.

Функції mean() та var() можна використовувати для обчислення середнього значення і дисперсії деякого набору чисел. Застосування sqrt() до результату роботи var() дасть стандартне відхилення. Або можна просто застосувати функцію sd():

```
>set.seed (3)

>y = rnorm (100)
>mean (y)

[1] 0.0110

>var (y)

[1] 0.7329
>sqrt (var (y))
[1] 0.8561

>sd (y)

[1] 0.8561
```

Графіки

Використання функції plot() є основним способом візуалізації даних в R. Наприклад, plot(x,y) створить графік залежності значень y від значень x. Є велика кількість додаткових опцій, які можна подати на функцію plot(). Наприклад, використання аргументу xlab призведе до появи заголовка осі X. Для отримання додаткової інформації про функції plot() введіть команду ?plot:

```
>x = rnorm (100)
>y = rnorm (100)
>plot (x, y)
>plot (x, y, xlab = “this is the x-axis”,
ylab =” this is the y-axis”, main =” Plot of X vs Y”)
```

Часто у нас буде виникати бажання зберегти побудований в R графік. Використовувана для цього команда буде залежати від типу файла, який ми хочемо створити. Наприклад, для створення PDF-файла ми використовуємо функцію pdf(), а для створення JPEG-файла – функцію jpeg().

```
>pdf (“Figure.pdf”)
>plot (x, y, col = “green”)
>dev.off () null device 1
```

Функція dev.off() повідомляє R, що ми завершили створення графіка. Як альтернативу ми можемо просто скопіювати вміст графічного вікна і вставити його в файл потрібного типу, на зразок документа Word.

Функцію seq() можна використовувати для створення послідовності чисел. Наприклад, seq(a, b) створює вектор з цілих чисел від a до b. Є багато опцій: наприклад, seq(0, 1, length = 10) створить послідовність з 10 чисел, які рівномірно розміщені на проміжку від 0 до 1. Команда 3:11 є скороченим варіантом seq(3, 11) для цілих чисел:

```
>x =seq (1, 10)
>x
[1] 1 2 3 4 5 6 7 8 9 10
>x = 1:10
>x
[1] 1 2 3 4 5 6 7 8 9 10
>x = seq (-pi, pi, length =50)
```

Тепер ми побудуємо кілька більш складних графіків. Функція contour() створює контурну діаграму для зображення тривимірних даних, які нагадують топографічну карту. Ця функція приймає три аргументи:

- вектор значень *x* (перший вимір);
- вектор значення *y* (другий вимір);
- матриця, елементи якої відповідають значенням *z* (третій вимір) для кожної пари координат (*x*, *y*).

Як і у випадку з функцією plot(), є багато вхідних параметрів, які можна використовувати для детального налаштування функції contour(). Щоб дізнатися про них детальніше, перегляньте довідковий файл, набравши ?controur:

```
>y = x
>f = outer (x, y, function (x, y) cos(y)/(1+x^2))
>contour (x, y, f)
>contour (x, y, f, nlevels=45, add=T)
>fa = (f-t(f))/2
>contour (x, y, fa, nlevels =15)
```

Функція image() працює так само, як і contour(), за винятком того, що вона створює кольорову діаграму, де колір залежить від значення. Така діаграма відома як теплова карта й іноді використовується для зображення температури в прогнозах погоди. Як альтернатива, для створення тривимірних графіків можна використовувати persp(). Аргументи theta та phi контролюють кути огляду графіка.

```
>image (x, y, fa)
>persp (x, y, fa)
>persp (x, y, fa, theta = 30)
>persp (x, y, fa, theta = 30, phi = 20)

>persp (x, y, fa, theta = 30, phi = 70)
>persp (x, y, fa, theta = 30, phi=40)
```

ЗАВДАННЯ ДО ЛАБОРАТОРНОЇ РОБОТИ

Завдання 1. Завдання стосується набору даних College, який можна знайти в файлі College.csv. Він містить ряд змінних для 777 різних університетів та коледжів у США. До цих змінних належать:

- Private: індикатор публічного / приватного;
- Apps: кількість отриманих додатків;
- Accept: кількість прийнятих заявників;
- Enroll: кількість нових студентів, які зареєстровані;
- Top10perc: нові учні з 10 % кращих класів середньої школи;
- Top25perc: нові учні з 25 % кращих класів середньої школи;
- F.Undergrad: кількість студентів денної форми навчання;
- P.Undergrad: кількість студентів-заочників;
- Outstate : навчання за межами штату;
- Room.Board: вартість проживання та харчування;
- Books: орієнтовна вартість книг;
- Personal: передбачувані особисті витрати;
- PhD: відсоток викладачів із ступенем доктора філософії;
- Terminal: відсоток викладачів із кінцевим ступенем;
- S.F.Ratio: співвідношення студент / викладацький склад;
- perc.alumni: відсоток випускників, які роблять пожертви;
- Expend: витрати на навчання на учня;
- Grad.Rate: рівень випускників.

Перш ніж читати дані в R, їх можна переглянути в Excel або текстовому редакторі.

1. Використовуйте функцію read.csv(), щоб прочитати дані в R. Викличте завантажений датасет college. Переконайтеся, що ви налаштували правильну директорію, в якій зберігаються дані.

2. Перегляньте дані за допомогою функції fix(). Ви повинні помітити, що перша колонка – це просто назва кожного університету. Ми не дуже хочемо, щоб цей стовпець R розглядав як дані. Але це може стати в нагоді потім. Спробуйте такі команди:

```
> rownames (college )=college [,1]
> fix(college)
```

Ви повинні побачити, що тепер є стовпець row.names із назвою кожного записаного університету. Це означає, що R дав у кожному рядку ім’я, що відповідає відповідному університету. R не намагатиметься виконати обчислення над іменами рядків, однак нам ще потрібно видалити перший стовпець у даних, де імена зберігаються. Спробуйте:

```
> college =college [,-1]
> fix(college)
```

Тепер ви повинні побачити, що перший стовпець даних – Private. ***Примітка:*** інший стовпець із позначкою row.names тепер з’являється перед колонкою Private. Однак це не стовпець даних, а радше ім’я, яке R дає кожному рядку.

3. Скористайтеся функцією summary(), щоб створити числове зведення змінних у наборі даних.

4. Використовуйте функцію pairs(), щоб створити матрицю діаграми розсіювання перших десяти стовпців або змінних даних. Згадайте, що ви можете посилатися на перші десять стовпців матриці A за допомогою команди A [,1:10].

5. Створіть нову якісну змінну Elite способом розбиття на класи змінної Top10perc. Ми будемо ділити університети на дві групи залежно від того, чи перевищує 50 % частка студентів, які входять у список 10 %, кращих учнів своїх класів у середній школі.

```
> Elite=rep("No",nrow(college ))
> Elite[college$Top10perc >50]=" Yes"
> Elite=as.factor(Elite)
> college =data.frame(college, Elite)
```

Використовуйте функцію summary(), щоб дізнатися, скільки є елітних університетів. Тепер використовуйте функцію plot() для створення діаграми розмахів по змінних Outstate проти Elite.

6. Використовуйте функцію hist(), і для кількох кількісних змінних побудуйте гістограми з різним числом класів. Вам може бути корисною команда par(mfrow=c(2,2)): вона розділить вікно друку на чотири сектори, щоб можна було побудувати чотири графіки одночасно. Зміна аргументів цієї функції розділить екран інакше.

7. Продовжте вивчати дані та надайте короткий підсумок того, що ви відкриваєте.

Завдання 2. У цій вправі використовується набір даних Auto, який вивчався в лабораторній роботі. Переконайтесь, що відсутні значення були видалені з даних.

1. Які з предикторів є кількісними, а які якісними?

2. Який діапазон кожного кількісного предиктора? Ви можете відповісти на це за допомогою функції range().

3. Яке середнє значення та стандартне відхилення кожного кількісного предиктора?

4. Тепер видаліть спостереження номер 10 та 85. Чому дорівнює розмах, середнє значення та стандартне відхилення кожного предиктора в підмножині даних, що залишилось?

5. Використовуючи повний набір даних, дослідіть предиктори графічно, за допомогою діаграм розсіювання або інших інструментів на ваш вибір. Створіть кілька графіків, які характерезують взаємозв'язок між предикторами.

6. Припустимо, що ми хочемо передбачити витрату бензину (mpg) на основі інших змінних. Чи передбачають ваші графіки, що якісь змінні можуть бути корисними для прогнозування mpg? Обґрунтуйте свою відповідь.

Завдання 3. Ця вправа стосується набору даних Boston.

1. Для початку завантажте набір даних Boston. Набір даних Boston є частиною бібліотеки MASS:

```
> library(MASS)
Тепер набір даних міститься в об'єкті Boston:
> Boston
Почитайте про набір даних:
> ? Boston
```

Скільки рядків у цьому наборі даних? Скільки колонок? Що являють собою рядки та стовпці?

2. Створіть кілька діаграм розсіювання предикторів (стовпців) цього набору даних. Опишіть результати.

3. Чи пов'язаний будь-який із предикторів із рівнем злочинності на душу населення? Якщо так, поясніть зв'язок.

4. Чи є якісь із передмість Бостона з особливо високим рівнем злочинності? Ставки податків (tax)? Співвідношення учень-вчитель (ptratio)? Прокоментуйте розмах значень кожного предиктора.

5. Через скільки передмість у цьому наборі даних протікає річка Чарльз?

6. Яке середнє співвідношення учень / вчитель у передмістях у цьому наборі даних?

7. Яке передмістя Бостона має найнижчу середню вартість будинків, які займають власники? Які значення інших предикторів для цього передмістя та як ці значення відрізняються від значень відповідних предикторів загалом? Прокоментуйте свої висновки.

8. Скільки передмість має середнє число кімнат у будинку (rm), більше семи? Більше восьми? Прокоментуйте передмістя, у яких число кімнат у будинку в середньому більше восьми.

КОНТРОЛЬНІ ПИТАННЯ

1. Як ви встановлюєте та завантажуєте пакети в R?
2. Які існують базові структури даних в R?
3. Що таке функція в R і як ви визначаєте власну функцію?
4. Як ви створюєте графіки в R?
5. Як ви використовуєте операції злиття та об'єднання даних в R?
6. Як ви працюєте з категоріальними змінними в R?
7. Які інструменти ви використовуєте для візуалізації в R?
8. Які плюси та мінуси ви бачите в R, порівняно з іншими мовами програмування для аналізу даних?

ЛАБОРАТОРНА РОБОТА № 3

Тема: проста лінійна регресія

Мета: опрацювати поняття проста лінійна регресивна модель, побудова регресивної моделі в електронних таблицях та пакеті R.

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Серед багаточисленних зв'язків між економічними показниками завжди можна виділити такий показник, вплив якого на результативну ознаку є основним, найбільш важливим. Щоб виміряти цей зв'язок кількісно, необхідно побудувати економетричну модель з двома змінними (просту модель). Загальний вигляд такої моделі:

$$Y = f(X, u),$$

де Y – залежна змінна (результативна ознака);

X – незалежна змінна (фактор);

u – стохастичний складник.

Аналітична форма цієї моделі може бути різною, залежно від економічної сутності зв'язків. Найбільш поширені форми залежностей:

$$Y = a_0 + (a_1 X);$$
$$Y = a_0 e^{a_1 X};$$
$$Y = a_0 X^{a_1};$$
$$Y = a_0 + \frac{a_1}{X},$$

де a_0, a_1 – невідомі параметри моделі.

Неважко переконатись, що наведені нелінійні форми залежностей за допомогою елементарних перетворень приводяться до лінійних. Якщо припустити, що економетрична модель з двома змінними є лінійною:

$$Y = a_0 + a_1 X + u,$$

в якій стохастичний складник (залишки) має нульове математичне сподівання та постійну дисперсію, то параметри моделі можна оцінити на основі звичайного методу найменших квадратів (1МНК).

В основі методу 1МНК лежить принцип мінімізації суми квадратів залишків моделі. Реалізація цього принципу дає можливість отримати систему нормальних рівнянь:

$$\begin{cases} na_0 + \sum_i x_i a_1 = \sum_i y_i \\ \sum_i x_i a_0 + \sum_i x_i^2 a_1 = \sum_i x_i y_i \end{cases}.$$

У цій системі n – кількість спостережень, $\sum_i x_i$, $\sum_i y_i$, $\sum_i x_i^2$, $\sum_i x_i y_i$ – величини, які можна розрахувати на основі вихідних спостережень над змінними Y і X.

Розв'язавши систему нормальних рівнянь, одержимо оцінки невідомих параметрів моделі $\hat{a}_0$ і $\hat{a}_1$:

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 X.$$

Достовірність побудованої економетричної моделі можна перевірити, використовуючи елементи дисперсійного аналізу. Насамперед треба розрахувати залишки моделі:

$$u_i = y_i - \hat{y}_i$$

та знайти їх дисперсію:

$$\sigma_u^2 = \frac{\sum\limits_i u_i^2}{n-m},$$

де m – кількість змінних моделі ($m = 2$);

$$S_{\hat{a}_j} = \sigma_u \sqrt{c_{jj}}\ .$$

Необхідно визначити стандартну помилку кожного параметра моделі. Коефіцієнт c_{jj} в цій формулі характеризує відповідний діагональний елемент матриці помилок (матриці, оберненої до матриці системи нормальних рівнянь).

На основі коефіцієнта детермінації:

$$R^2 = \frac{\sigma_Y^2 - \sigma_U^2}{\sigma_Y^2}$$

можна зробити висновок про ступінь значущості вимірюваного зв'язку на основі економетричної моделі:

$$R^2 \in]0,1[\ .$$

Оскільки коефіцієнт детермінації R^2 характеризує, якою мірою варіація залежної змінної визначається варіацією незалежної змінної, то чим ближче R^2 до одиниці, тим суттєвішим є зв'язок між цими змінними.

Коефіцієнт кореляції $R = \sqrt{R^2}$ характеризує тісноту зв'язку між змінними моделі. Він може знаходитись на множині $R \in]-1,1[$. Чим ближче R до одиниці по модулю, тим тіснішим є зв'язок. Від'ємний знак свідчить про обернений зв'язок, додатний – про прямий.

Якщо прийняти відповідну гіпотезу про закон розподілу залишків економетричної моделі, то її параметри можна оцінити на основі методу максимальної правдоподібності.

Нехай залишки моделі розподіляються за нормальним законом, тоді функція правдоподібності запишеться так:

$$F = \frac{1}{(\tilde{\sigma}_u^2 2\pi)^{n/2}} \exp\left[-\frac{1}{2\tilde{\sigma}_u^2} \sum_{i=1}^{n} (y_i - \tilde{a}_0 - \tilde{a}_1 x_i)^2\right]$$

і

$$\ln F = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \tilde{\sigma}_u^2 - \frac{1}{2\tilde{\sigma}_u^2} \sum_{i=1}^{n} (y_i - \tilde{a}_0 - \tilde{a}_1 x_i)^2\ .$$

Продифереціюємо цю функцію за невідомими параметрами $\tilde{a}_0$, $\tilde{a}_1$ і $\tilde{\sigma}_u^2$ і, прирівнявши похідні до нуля, отримаємо систему рівнянь:

$$\begin{cases} n\tilde{a}_0 + \tilde{a}_1 \sum_i x_i = \sum_i y_i \\ \sum_i x_i \tilde{a}_0 + \tilde{a}_1 \sum_i x_i^2 = \sum_i x_i y_i \\ \frac{1}{n} \sum_i (y_i - \tilde{a}_0 - \tilde{a}_1 x_i)^2 = \sigma_U^2. \end{cases}$$

Підставимо в цю систему величини $\sum_i x_i$, $\sum_i y_i$, $\sum_i x_i^2$, $\sum_i x_i y_i$, які розраховуються на основі вихідних даних, і розв'яжемо її відносно параметрів $\tilde{a}_0$, $\tilde{a}_1$ і $\tilde{\sigma}_u^2$. У результаті отримаємо оцінки параметрів моделі $\tilde{a}_0$ і $\tilde{a}_1$, а також оцінку дисперсії залишків.

Економетрична модель з двома змінними: побудова та аналіз

Приклад 3.1

На основі даних про роздрібний товарообіг і доходи населення побудувати економетричну модель роздрібного товарообігу. Дати загальну характеристику достовірності моделі та зробити висновки.

Вихідні дані та елементарні перетворення цих даних для побудови моделі наведені в табл. 3.1.

Таблиця 3.1 – Дані для побудови моделі.

№ з/п	Y	X	X^2	XY	$\hat{Y}$	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(X-\bar{X})*(Y-\bar{Y})$	$u=Y-\hat{Y}$	u^2	$(Y-\bar{Y})^2$
1	17	18	324	306	16.67	–6.5	–5	42.25	32.5	0.33	0.1089	25
2	18	20	400	360	18.31	–4.5	–4	20.25	18.0	–0.31	0.0961	16
3	19	21	441	399	19.31	–3.5	–3	12.25	10.5	–0.13	0.0169	9
5	21	24	576	504	21.59	–0.5	–1	0.25	0.5	–0.59	0.3481	1
6	23	25	625	575	22.41	0.5	1	0.25	0.5	0.59	0.3481	1
7	24	27	729	648	24.05	2.5	2	6.25	5.0	–0.05	0.0125	4
8	25	28	784	700	24.87	3.5	3	12.25	10.5	0.13	0.0169	9
9	26	29	841	754	25.69	4.5	4	20.25	18.0	0.31	0.0961	16
10	27	31	961	837	27.33	6.5	5	42.25	32.5	–0.33	0.1089	25
10 $\sum_{i=1}$	220	245	6 165	5 523	-----	----	--	162.5	133	----	1.145	110

Розв'язання

1. Ідентифікуємо змінні:

Y – роздрібний товарообіг (залежна змінна);

X – доходи населення (незалежна змінна).

2. Нехай специфікація моделі $Y = f(X,u)$ визначається лінійною функцією; вона має такий вигляд:

$$Y = a_0 + a_1 X + u,$$

де a_0, a_1 – параметри моделі;

u – стохастичний складник, залишки.

3. Оцінимо параметри моделі $\hat{Y} = \hat{a}_0 + \hat{a}_1 X$ за методом 1МНК. Для цього запишемо систему нормальних рівнянь:

$$\begin{cases} n\hat{a}_0 + \sum_i x_i \hat{a}_1 = \sum_i y_i & (i = \overline{1,n}); \\ \sum_i x_i \hat{a}_0 + \sum_i x_i^2 \hat{a} = \sum_i x_i y_i & (i = \overline{1,n}); \end{cases}$$

$n = 10$ – кількість спостережень.

Для знаходження параметрів моделі використаємо надбудову електронного процесора Excel «Аналіз даних». На рис. 3.1 показана роздруківка роботи надбудови.

ВИВІД ПІДСУМКІВ								
Регресійна статистика								
Множинний R	0,994784							
R-квадрат	0,989594							
Нормований R-квадрат	0,988294							
Стандартна помилк	0,378255							
Спостереження	10							
Дисперсійний аналіз								
	df	SS	MS	F	начимість F			
Регресія	1	108,8554	108,8554	760,8172	3,219E-09			
Залишок	8	1,144615	0,143077					
Підсумок	9	110						
	Коефіцієнти	Стандартна помилка	t-статистика	P-Значення	Нижні 95%	Верхні 95%	Нижні 95,0%	Верхні 95,0%
Y-перетин	1,947692	0,736758	2,6436	0,029548	0,248726	3,646658	0,248726	3,646658
Зміна X 1	0,818462	0,029673	27,58292	3,219E-09	0,750036	0,886887	0,750036	0,886887

Рис. 3.1. Роздруківка роботи надбудови електронного процесора Excel «Аналіз даних»

Отже, економетрична модель запишеться так:

$$Y = 1{,}94 + 0{,}81_x.$$

4. Знайшовши відхилення кожної змінної від своєї середньої арифметичної, розрахуємо параметри моделі альтернативним способом:

$$\alpha_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{133}{162{,}5} = 0{,}81;$$

$$\alpha_0 = y - \alpha_1 x = 22 - 0{,}81 * 24{,}5 = 22 - 20{,}06 = 1{,}94.$$

5. Визначимо коефіцієнти детермінації та кореляції:

$$R^2 \approx 0{,}98;$$
$$R = 0{,}994.$$

Оскільки коефіцієнт детермінації $R^2 \approx 0{,}98$, це свідчить, що варіація обсягу роздрібного товарообігу на **98 %** визначається варіацією доходів населення. Коефіцієнт кореляції $R = 0{,}994$ характеризує тісний зв'язок між цими соціально-економічними показниками. Величини R^2 і R для парної економетричної моделі свідчать про її достовірність, якщо вони наближаються до одиниці.

6. Визначимо за роздруківкою адекватність моделі. Модель є адекватною, тому що у стовпчику «Значимість F» число наближається до нуля.

7. Визначимо стандартні помилки оцінок параметрів моделі, враховуючи дисперсію залишків:

$$S_{a_0} = 0{,}736;$$
$$S_{a_1} = 0{,}029.$$

Порівняємо стандартні помилки оцінок параметрів моделі з величиною цих оцінок. У результаті визначимо, що стандартна помилка оцінки параметра $\hat{a}_1$ становить **3,4 %** абсолютного значення цієї оцінки **(0,81),** що свідчить про незміщеність цієї оцінки параметра моделі. Стандартна помилка оцінки параметра $\hat{a}_0$ становить **38 %** абсолютного значення цієї оцінки **(1,94),** а це означає, що цей параметр може мати зміщення, яке зумовлюється невеликою сукупністю спостережень ($\boldsymbol{n = 10}$).

Висновок. Регресивна модель $\hat{Y} = 1{,}91 + 0{,}82X$ кількісно описує зв'язок роздрібного товарообігу і доходів населення.

Параметр $\hat{a}_1 = 0{,}82$ характеризує граничну величину витрат на купівлю товарів у роздрібній торгівлі, коли дохід збільшується на одиницю, тобто у разі збільшення доходів на одиницю обсяг роздрібного товарообігу зростає на 0,82 одиниці.

ЗАВДАННЯ ДО ЛАБОРАТОРНОЇ РОБОТИ

Завдання 1. Завдання стосується використання простої лінійної регресії на наборі даних Auto.

1. Використовуйте функцію lm(), щоб виконати просту лінійну регресію, де mpg як відгук і horsepower як предиктор. Використайте функцію summary() для виводу результатів. Прокоментуйте результат. Наприклад:

Чи існує зв'язок між предиктором і відгуком?

Наскільки сильний зв'язок між предиктором і відгуком?

Цей зв'язок між предиктором і відгуком позитивний чи негативний?

Чому дорівнює модельне значення mpg за hoursepower із індексом $90 + n^2$*? (де* n *= номер студента в списку групи (не потоку)).*

Які відповідні 95 -відсоткові довірчі інтервали для регресійної прямої і для передбаченого значення?

Використовувати функцію predict() зі значенням interval = «confidence» та interval = «prediction» відповідно.

2. Побудуйте графік залежності відгука і предиктора. Використовуйте функцію abline(), щоб відобразити лінію регресії методом найменших квадратів.

3. Використовуйте функцію plot() для створення діагностичних графіків для моделі. Прокоментуйте будь-які проблеми, які ви бачите у цій моделі.

Завдання 2. Виконайте завдання 1, але з датасетом Abalone. Відгук берете за Rings, а предиктор за варіантом із таблиці 3.2.

Таблиця 3.2 – Варіанти предикторів

Номер варіанта *n*	**Предиктор**
1	Length
2	Diameter
3	Height (додатково потрібно позбутися викидів)
4	Whole weight
5	Shucked weight
6	Viscera weight
7	Shell weight
8	Length
9	Diameter
10	Height
11	Whole weight
12	Shucked weight

КОНТРОЛЬНІ ПИТАННЯ

1. Сформулюйте особливості лінійної регресивної моделі.

2. Які вимоги висуваються до обсягу спостережень, необхідного для побудови регресійної моделі?

3. Поясніть суть методу найменших квадратів для оцінювання параметрів лінійних регресивних моделей.

4. Запишіть різні форми системи нормальних рівнянь для множинної лінійної моделі. Якими методами може бути розв’язана система нормальних рівнянь у цьому випадку?

5. У чому полягає сутність понять «незміщеність», «обґрунтованість» і «ефективність оцінок»? Які гіпотези повинні задовольняти відхилення в моделі, щоб оцінки параметрів моделі, отримані за допомогою МНК, мали властивості незміщеності, обґрунтованості й ефективності?

6. У зв’язку з чим необхідно перевіряти статистичну значущість оцінок параметрів моделі?

7. У чому полягає суть критерію Стьюдента? Як визначається статистична значущість оцінок параметрів моделі?

8. Як визначаються довірчі інтервали для оцінок параметрів?

9. Що таке адекватність моделі? Назвіть методи визначення адекватності моделі.

10. У чому полягає суть коефіцієнта лінійної кореляції? Якими методами можна його розрахувати?

11. Як здійснюється розрахунок прогнозних значень за лінійною регресивною моделлю?

12. Чим розрізняються рівняння регресії в натуральному та стандартизованому виглядах?

13. Як створити набір даних для простої лінійної регресії в R?

14. Яка функція використовується для побудови простої лінійної регресії в R?

15. Як інтерпретувати коефіцієнти моделі простої лінійної регресії в R?

16. Як можна оцінити якість моделі простої лінійної регресії?

17. Які діагностичні графіки використовуються для перевірки припущень лінійної регресії в R?

18. Як отримати прогнозовані значення з використанням моделі простої лінійної регресії в R?

19. Як використовувати візуалізацію для ілюстрації результатів простої лінійної регресії в R?

ЛАБОРАТОРНА РОБОТА № 4

Тема: множинна лінійна регресія.

Мета: опрацювати поняття «множинна лінійна регресивна модель», «коефіцієнти еластичності», «побудова регресивної моделі в електронних таблицях та пакеті R».

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Для того, щоб кількісно описати зв'язок між кількома або багатьма змінними, одна з яких є залежною, інші – незалежними змінними, необхідно розглянути лінійну економетричну модель, яка базується на регресійному аналізі.

У загальному вигляді цю модель можна записати так:

$$Y=f(X_1,X_2,...,X_m,u),$$

де Y – залежна змінна;

X_j, $(j=\overline{1,m})$ – незалежні змінні;

u – стохастичний складник.

Залежна змінна Y називається також пояснюваною, ендогенною змінною, незалежні змінні X_j – пояснювальними, предетермінованими, екзогенними змінними.

Аналітична форма загальної лінійної економетричної моделі:

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_mX_m + u,$$

де a_j $(j=\overline{0,m})$ – параметри моделі.

У матричній формі економетрична модель має такий вигляд:

$$Y = XA + u,$$

X – матриця незалежних змінних;

A – вектор оцінок параметрів моделі;

u – вектор залишків.

Щоб оцінити параметри моделі на основі методу 1МНК, необхідно дотримуватися таких передумов (гіпотез):

1) математичне сподівання залишків має дорівнювати нулю, тобто:

$$\mathrm{M}(u) = 0;$$

2) значення вектора залишків u незалежні між собою і мають постійну дисперсію:

$$\mathrm{M}(uu') = \sigma^2\mathrm{E};$$

3) незалежні змінні моделі не зв'язані із залишками, тобто:

$$\mathrm{M}(X'u) = 0;$$

4) незалежні змінні моделі створюють лінійно-незалежну систему векторів, тобто:

$$\begin{cases} (X_k'\,X_j) = 0, & k \neq j; \qquad k = \overline{1,m} \\ (X_k'\,X_j) = 1, & k = j; \qquad j = \overline{1,m}\,. \end{cases}$$

Оператор оцінювання параметрів моделі на основі 1МНК:

$$A = (X'X)^{-1} X'Y.$$

Неважко довести, що оцінки $\hat{A}$, які можна отримати на основі оператора оцінювання 1МНК, мінімізують суму квадратів залишків u. Водночас значення вектора $\hat{A}$ є розв'язком нормальної системи рівнянь:

$$(X'X)\hat{A} = X'Y.$$

Якщо незалежні змінні в матриці X взяті як відхилення кожного значення від своєї середньої, то матрицю $X'X$ називають матрицею моментів. Числа, що стоять на її головній діагоналі, характеризують величину дисперсій незалежних змінних, інші елементи відповідають взаємним коваріаціям.

Оцінки параметрів загальної економетричної моделі повинні мати такі *властивості:*

1) незміщеності;
2) обґрунтованості;
3) ефективності;
4) інваріантності.

Оцінка параметра моделі буде *незміщеною,* коли дотримується рівність:

$$M(\hat{A}) = A.$$

Якщо ця рівність не дотримується, то різниця $M(\hat{A}) - A = Q$ називається зміщенням оцінки.

Оцінка параметра моделі буде *обґрунтованою,* якщо за заданої малої величини $\varepsilon > 0$ справедливе відношення:

$$\lim_{n\to\infty} P\left\{ \left|\hat{A} - A\right| < \varepsilon \right\} = 1 .$$

Оцінки $\hat{A}$ параметрів A називаються *ефективними,* коли вони мають найменшу дисперсію.

Якщо функція $g(\hat{A})$ відповідає функції g(A), то оцінки $\hat{A}$ параметрів A є інваріантними.

Приклад

Побудувати економетричну модель, яка характеризує залежність між витратами на харчування, загальними витратами та складом сім'ї. Проаналізувати зв'язок, визначений на основі побудованої моделі.

Розв'язання

1. Ідентифікуємо змінні моделі:

Y – витрати на харчування (залежна змінна);

X_1 – загальні витрати (незалежна змінна);

X_2 – розмір сім'ї (незалежна змінна);

u – залишки (стохастичного складника).

Загальний вигляд моделі:

$$Y = f(X_1, X_2, u) .$$

2. Специфікуємо модель, тобто в цьому випадку визначимо її аналітичну форму:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + u;$$

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2 .$$

3. Оцінимо параметри моделі на основі методу 1МНК, попередньо висунувши гіпотезу, що всі чотири передумови для його застосування дотримані.

Отже, $\hat{a}_0 = 8{,}347; \hat{a}_1 = 0{,}167; \hat{a}_2 = 8{,}175$. Звідси, регресивна модель має вигляд:

$$Y = 8{,}347 + 0{,}167\ X_1 + 8{,}175\ X_2 .$$

4. Випишемо стандартні помилки оцінок параметрів:

$$S_{a_0} = 6{,}559;$$

$$S_{a_1} = 0{,}022;$$

$$S_{a_2} = 1{,}598.$$

Порівняємо стандартні помилки оцінок параметрів моделі з величиною оцінки. Так, співвідношення стандартної помилки й абсолютного значення параметра $\hat{a}_0$ становить **56 %,** параметра $\hat{a}_1$ – **10,6 %,** параметра $\hat{a}_2$ – **20,4 %.** Перше й третє співвідношення свідчать про те, що оцінки параметрів моделі $\hat{a}_0$ і $\hat{a}_2$ можуть мати зміщення, а друге співвідношення підтверджує незміщеність оцінки параметра $\hat{a}_1$.

5. Множинний коефіцієнт кореляції дорівнює 0,974. Коефіцієнт детермінації становить 95 %. Отже, зміну витрат на харчування на 95 % визначає зміна загальних витрат та складу сім'ї.

6. Модель є адекватною, тому що значущість F наближається до нуля.

ВИВІД ПІДСУМКІВ

Регресійна статистика	
Множинний R	0,974838
R-квадрат	0,950308
Нормований R-квадрат	0,942664
Стандартна помилка	11,81472
Спостереження	16

Дисперсійний аналіз

	df	SS	MS	F	Значимість F
Регресія	2	34703,36	17351,68	124,3067	3,35615E-09
Залишок	13	1814,639	139,5876		
Підсумок	15	36518			

	Коефіцієнти	Стандартна помилка	t-статистика	P-Значення	Нижні 95%	Верхні 95%	Нижні 95,0%	Верхні 95,0%
Y-перетин	8,3476	6,559893	1,272521	0,225477	-5,824186602	22,51939	-5,82419	22.51939
Зміна X 1	0,16753	0,022263	7,52507	4,3416E-0(	0,119433633	0,215626	0,119434	0,215626
Зміна X 2	8,175258	1,598968	5,112833	0,000199	4,720896779	11,62962	4,720897	11,62962

Рисунок 4.1. Роздруківка роботи надбудови електронного процесора Excel «Аналіз даних»

7. Дамо змістовне тлумачення параметрів моделі.

Оцінка параметра $\hat{a}_1$ характеризує граничну зміну величини витрат на харчування, залежно від зміни загальних затрат на одиницю. Тобто якщо загальні витрати сім'ї зростуть на одиницю, то витрати на харчування в них збільшаться на **0,18** одиниці за незмінного складу сім'ї.

Оцінка параметра $\hat{a}_2$ характеризує граничне зростання витрат на харчування за умови збільшення сім'ї на одного члена. Так, якщо склад сім'ї збагатиться ще на одного члена, то витрати на харчування зростуть на **6,854** одиниці за незмінної величини доходу.

8. Знайдемо стандартизовані параметри регресії.

9. Знайдемо коефіцієнти еластичності.

ЗАВДАННЯ ДО ЛАБОРАТОРНОЇ РОБОТИ

Завдання 1. Побудувати множинну лінійну регресію на наборі даних Auto.

1. Створіть матрицю діаграми розсіювання, яка включає всі змінні у цьому наборі даних.

2. Обчисліть матрицю кореляції між змінними за допомогою функції cor(). Вам потрібно буде виключити якісні змінні.

3. Використовуйте функцію lm(), щоб виконати множинну лінійну регресію з mpg як відгук та всіма іншими змінними, крім name як предиктори. Для виводу результатів використовуйте функцію summary(). Прокоментуйте результат. Наприклад:

Чи існує зв'язок між предикторами та відгуком?

Які предиктори є статистично значущими відносно відгуку?

Що означає коефіцієнт для змінної year?

4. Використовуйте функцію plot() для створення діагностичних графіків для цієї лінійної регресійної моделі. Прокоментуйте будь-які проблеми з цією моделлю. Чи свідчать залишкові графіки про якісь надзвичайно великі викиди?

Завдання 2. Це завдання передбачає використання множинної лінійної регресії на наборі даних Abalone.

1. Створіть матрицю діаграми розсіювання, яка включає Rings та змінні за варіантом із таблиці 2 у цьому наборі даних.

2. Обчисліть матрицю кореляції між змінними за допомогою функції cor().

3. Використовуйте функцію lm(), щоб виконати множинну лінійну регресію з Rings як відгук та змінними по варіанту як предиктори. Для виводу результатів використовуйте функцію summary(). Прокоментуйте результат. Наприклад:

Чи існує зв'язок між предикторами та відгуком?

Які предиктори є статистично значущими відносно відгуку?

Що означає коефіцієнт для кожної змінної?

4. Використовуйте функцію plot() для створення діагностичних графіків для цієї лінійної регресійної моделі. Прокоментуйте будь-які проблеми з цією моделлю. Чи свідчать залишкові графіки про якісь надзвичайно великі викиди?

Таблиця 4.1 – Варіанти предикторів

Номер варіанта n	**Предиктори**
1	Length Diameter Height (додатково потрібно позбутися викидів)
2	Diameter Height (додатково потрібно позбутися викидів) Whole weight
3	Height (додатково потрібно позбутися викидів) Whole weight Shucked weight
4	Whole weight Shucked weight Viscera weight
5	Shucked weight Viscera weight Shell weight
6	Viscera weight Shell weight Length
7	Shell weight Length Diameter
8	Length Diameter Whole weight
9	Diameter Height (додатково потрібно позбутися викидів) Viscera weight
10	Height Whole weight Viscera weight
11	Length Whole weight Shucked weight
12	Diameter Whole weight Shell weight

КОНТРОЛЬНІ ПИТАННЯ

1. Сформулюйте особливості нелінійної регресивної моделі.

2. Які вимоги висуваються до обсягу спостережень, необхідного для побудови багатофакторної регресійної моделі?

3. Поясніть суть методу найменших квадратів для оцінювання параметрів множинних регресивих моделей.

4. Запишіть різні форми системи нормальних рівнянь для множинної лінійної моделі. Якими методами може бути розв’язана система нормальних рівнянь у цьому випадку?

5. Як визначаються довірчі інтервали для оцінок параметрів множинної регресивної моделі?

6. Що таке адекватність моделі? Назвіть методи визначення адекватності моделі.

7. У чому полягає сутність коефіцієнта множинної кореляції? Якими методами можна його розрахувати?

8. Як здійснюється розрахунок прогнозних значень за множинною регресивною моделлю?

9. Чим розрізняються рівняння регресії в натуральному та стандартизованому виглядах?

10. Яка функція в R використовується для побудови множинної лінійної регресії?

11. Як підготувати дані для аналізу множинної лінійної регресії в R?

12. Як перевірити результати моделі множинної лінійної регресії?

13. Як візуалізувати результати множинної лінійної регресії в R?

14. Які статистичні показники можна використовувати для оцінки якості моделі?

ЛАБОРАТОРНА РОБОТА № 5

Тема: перетворення змінних та ефекти взаємодії лінійної регресії.

Мета: навчитись виконувати перетворення змінних для зведення нелінійної залежності до лінійної та розраховувати ефекти взаємодії лінійної регресії.

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Для вибору й обґрунтування типу кривої регресії немає універсального методу. Однобічна стохастична залежність між явищами може бути описана, наприклад, за допомогою поліноміальної регресії:

$$\hat{y} = b_0 + b_1 \mathrm{x} + b_2 x^2 + b_0 x^3 + \cdots, \quad (1)$$

або за допомогою гіперболічної регресії:

$$\widehat{y^*} = b_0^* + b_1^* \frac{1}{x}. \quad (2)$$

Використовуються також степенева, показникова, логарифмічна і тригонометрична функція. Підбір функції регресії має виконуватись із застосуванням теорії тієї конкретної науки, на базі якої виникає задача виміру зв'язку між явищами. Найчастіше використовуються сімейства кривих, рівняння яких виражаються багаточленами цілих позитивних ступенів (поліноми виду (1)).

Поліном першого ступеня (пряма лінія) не має вигинів. За допомогою полінома другого ступеня можна передати одну точку повороту функції. Поліном третього ступеня відбиває двох точок повороту функції.

Про характер залежності між економічними явищами часто роблять висновок за зовнішнім виглядом емпіричного графіка регресії. Однак за умови малого числа спостережень цей шлях приводить до незадовільних результатів, тому що різкі зигзаґи емпіричної (ламаної) лінії регресії ускладнюють виявлення закономірностей. У кожному випадку варто перевіряти можливість застосування лінійної регресії хоча б на обмеженій ділянці зміни змінних. Необхідно звертати увагу на те, щоб оцінки регресії вироблялися з достатнім ступенем надійності. Інформацію про це нам дає коефіцієнт детермінації.

Ми розрізняємо два класи нелінійних регресій. **До першого класу** належать регресії, нелінійні щодо включених в аналіз пояснювальних змінних Xk, але лінійні за невідомими параметрами регресії $b_k (k = 1,2, \dots, p)$, що підлягають оцінюванню. Тому утворюючи цей клас нелінійні регресії називають також квазілінійними регресіями. Їх перевага полягає в тому, що для них можливе безпосереднє застосування методу найменших квадратів, а отже, залишаються в силі усі вихідні передумови лінійного регресійного аналізу і властивості МНК-оцінок параметрів регресії (незміщеність, гомоскедастичність та ін.). Використовуються ті ж самі критерії значущості, аналогічно будуються довірчі інтервали і довірчі зони.

Другий клас регресій характеризується нелінійністю за оцінюваними параметрами. Цей клас регресій зустрічається доволі часто під час досліджень економічних явищ, однак він має суттєвий недолік – не допускає застосування звичайного методу найменших квадратів. Для рішення системи нелінійних рівнянь, що виходить під час цього, залучають ітераційні методи або вдаються до апроксимації

параметрів шуканої залежності. Широко використовується також лінійне перетворення функції регресії, що дає змогу застосовувати до перетворених параметрів статистичні критерії лінійної регресії. Суворої теорії нелінійної регресії поки немає.

Квазілінійна регресія

Розглянемо спочатку ***просту квазілінійну регресію.*** Нехай залежність між двома явищами (y і x) представлена у вигляді параболи другого порядку (цілої раціональної функції другого ступеня:

$$\hat{y} = b_0 + b_1 \mathrm{x} + b_2 x^2, \qquad (3)$$

де b_0 – постійна, яка вирівнює та яка відповідає точці перетинання кривої регресії з віссю y;

b_1 та b_2 – параметри регресії, що характеризують залежність перемінної y від змінної x.

Функція регресії (3) лінійна щодо параметрів і нелінійна щодо пояснювальних змінних x (квадратний тричлен). Отже, ми маємо типову функцію квазілінійної регресії.

Для оцінки параметрів (3) методом найменших квадратів потрібно виходити зі співвідношення:

$$\hat{y} = b_0 + b_1 x_i + b_2 x_i^2 + \widehat{u_i}, \ \mathrm{i} = 1, 2, \ldots, \mathrm{n}, \qquad (4)$$

де u – похибка.

Прирівнявши до нуля частинні похідні від суми квадратів похибок за кожним із параметрів b_0, b_1, b_2, одержимо після деяких перетворень такі нормальні рівняння:

$$\sum_i y_i = n b_0 + b_1 \sum_i x_i + b_2 \sum_i x_i^2 ,$$

$$\sum_i x_i y_i = b_0 \sum_i x_i + b_1 \sum_i x_i^2 + b_2 \sum_i x_i^3 ,$$

$$\sum_i x_i^2 y_i = b_0 \sum_i x_i^2 + b_1 \sum_i x_i^3 + b_2 \sum_i x_i^4 .$$

Підставляючи в (3) обчислені значення b_0, b_1, b_2, ми знайдемо оцінку функції регресії. Після перевірки значущості оцінок параметрів регресії за прийнятної величини коефіцієнта детермінації можна визначити розрахункові значення регресії для аналізу залежності між економічними явищами.

Нехай, з огляду на логічні міркування, для опису залежності використовується гіперболічна форма зв’язку:

$$\widehat{y^*} = b_0^* + b_1^* \frac{1}{x}. \qquad (5)$$

Застосовуючи метод найменших квадратів до (5), знову одержимо систему нормальних рівнянь. Вирішуючи її, знаходимо b_0 і b_1:

$$b_0^* = \frac{\sum_i y_{i\,\sum_i \frac{1}{x_i^2}} - \sum_i \frac{y_i}{x_i} \sum_i \frac{1}{x_i}}{n\sum_i \frac{1}{x_i^2} - \left(\sum_i \frac{1}{x_i}\right)^2};$$
$$b_1^* = \frac{n\sum_i \frac{y_i}{x_i} - \sum_i \frac{1}{x_i} \sum_i y_i}{n\sum_i \frac{1}{x_i^2} - \left(\sum_i \frac{1}{x_i}\right)^2}. \quad (6)$$

За формулами (6) ми обчислюємо оцінки параметрів гіперболічного рівняння регресії.

Розглянемо тепер ***у загальному вигляді квазілінійну регресію, тобто функцію, нелінійну за пояснювальними змінними, але лінійну за оцінюваними параметрами:***

$$y = b_0 + b_1 F_1(x) + b_2 F_2(x) + \cdots + b_p F_p(x), \quad (7)$$

де $F_1(x)$, $F_2(x)$, … – функції від пояснювальних змінних x. Вони не містять інших параметрів. Так, наприклад, це можуть бути функції виду $F_1(x) = \log x$ чи $F_1(x) = 1/x$, але не такі, як $F_1(x) = \log(x - k)$ чи $F_2(x) = 1/xk$.

Застосовуючи метод найменших квадратів до (7), одержимо систему нормальних рівнянь:

$$\begin{gathered} \sum_i y_i = nb_0 + b_{1\,\sum_i F_1(x)} + b_{2\,\sum_i F_a(x)} + \cdots, \\ \sum_i y_i F_1(x) = b_{0\,\sum_i F_1(x)} + b_{1\,\sum_i F_1^2(x)} + b_{2\,\sum_i F_1(x) F_2(x)} + \cdots, \\ \sum_i y_i F_2(x) = b_{0\,\sum_i F_2(x)} + b_{1\,\sum_i F_1(x) F_2(x)} + b_{2\,\sum_i F_2^2(x)} + \cdots. \end{gathered} \quad (8)$$

З (8) можна вивести правило складання нормальних рівнянь. З огляду на те, що окремі значення сумуються, рівняння (8-1) будується аналогічно до регресії (7-1). Нормальні рівняння (8-1, 2, 3) та ін. виходять, якщо функцію регресії (7) помножити відповідно на $F_i(x)$ і $F_y(x)$, а потім додати. Це правило можна сформулювати, розглядаючи також нормальні рівняння для простої і множинної регресії.

Перш ніж перейти до прикладу, складемо зведення квазілінійних функцій, застосовуваних в економіці (див. табл. 1).

Розв’язуючи систему нормальних рівнянь, ми знаходимо параметри регресії. Вкажемо ще один спосіб представлення квазілінійних функцій у вигляді лінійної множинної регресії. У цьому випадку часто йдеться про функціональну регресію. Так, наприклад, зробивши в поліномі другого ступеня:

$$\widehat{y^*} = b_0^* + b_1^* x + b_2^* x^2, \quad (9)$$

наступну заміну:

$$\begin{gathered} x = x_1;\ x^2 = x_2; \\ b_0^* = b_0;\ b_1^* = b_1;\ b_2^* = b_2, \end{gathered} \quad (10)$$

можна записати його у вигляді:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2. \quad (11)$$

Ми одержали форму запису лінійної множинної регресії. Отже, формули для визначення коефіцієнтів множинної регресії b_0 і і b_0 придатні також із використанням (9, 10, 11) для знаходження параметрів нелінійної простої регресії.

Регресійні залежності, нелінійні за оцінюваними параметрами (2 клас функцій зростання), в економіці зустрічаються доволі часто. Використання цього класу регресій пов'язано з обчислювальними труднощами, тому що зазначені регресії не допускають безпосереднього застосування звичайного методу найменших квадратів. Для того, щоб зробити це можливим, вихідні дані піддають перетворенням, головне призначення яких у лінеаризації розглянутих залежностей за оцінюваними параметрами. Так, наприклад, у спосіб логарифмічного перетворення можна перейти від залежності показового типу до лінійного:

$$\hat{y} = ab^x, \qquad \log \hat{y} = \log a + x \log b\,. \tag{12}$$

Зробивши в (12) заміну $\log y = Z$, $\log a = A$ і $\log b = B$, одержимо:

$$\mathrm{Z} = \mathrm{A} + \mathrm{Bx}. \tag{13}$$

До рівняння (13) застосовуємо метод найменших квадратів.

Для визначення залежності виду (12) треба виконати логарифмічне перетворення змінної y, тобто прологарифмувати емпіричні значення y_i, $Z_i = \log y_i$.

Отже, деякі функції за допомогою перетворення змінних піддаються лінеаризації за своїми параметрами. Параметри регресії вихідних функцій знаходять шляхом зворотних перетворень. Наприклад, якщо вихідна функція є показниковою чи статичною із дробовим показником, то оцінки параметрів цих регресій знаходять через потенціювання параметрів лінеаризованих залежностей. Лінеаризація зв'язків дає можливість застосовувати для перебування оцінок параметрів метод найменших квадратів. Але отримані оцінки параметрів вихідних функцій можуть не мати властивості МНК-оцінок. Розроблено способи уточнення цих оцінок.

Для наочності в економіці нелінійні функції другого класу, що найбільш часто зустрічаються, представлені у табл. 3. Особливо значну роль вони відіграють підчас вивчення попиту. З наведених у таблиці функцій найбільше ускладнення під час їх визначення викликають оцінки параметрів a логістичної функції і функції Гомперца, параметра з функції Джонсона і параметра b функції Торнквиста 2-го і 3-го типів. Тому що параметр a вказує рівень насичення, і зазвичай він заздалегідь установлюється з огляду на логіко-економічні розуміння. Є також чисельні методи, за допомогою яких можна обчислити це значення. Інший спосіб полягає у визначенні рівня насичення за допомогою функції Торнквиста 1-го типу і підстановки цього значення в логістичну функцію. Але в будь-якому випадку ми повинні зважати на економічний аналіз явища.

Нелінійна кореляція

Якщо між явищами, що досліджуються, існують нелінійні відносини, то також, як і у випадку лінійного зв'язку, визначається щільність та сила залежності. Коефіцієнт кореляції для цього використовувати не можна, бо він має таку форму для розрахунку, що передбачає лінійну залежність між показниками, і він не буде точно відображувати інтенсивність зв'язку. Таким показником сили зв'язку слугує індекс кореляції.

Розглянемо вимірювання інтенсивності нелінійного зв'язку, коли між двома явищами об'єктивно наявна залежність виражається за допомогою квазілінійної функції:

$$R_{yx} = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}},$$

поділивши чисельник та знаменник на (n) чи $(n-1)$, отримаємо:

$$R_{yx} = \sqrt{\frac{s_{\hat{y}}^2}{s_y^2}} = \sqrt{1 - \frac{s_e^2}{s_y^2}}.$$

Після низки перетворень можна визначити ще таку формулу:

$$R_{yx} = \sqrt{\frac{n\sum \hat{y}_i^2 - (\sum y_i)^2}{n\sum {y_i}^2 - (\sum y_i)^2}}.$$

Індекс кореляції приймає значення $0 \le R_{yx} \le 1$.

Чим більше значення індексу кореляції, тим сильніший зв'язок. Індекс кореляції дорівнює кореню квадратного від коефіцієнта детермінації. Тому чим більше наближається індекс кореляції до 1, тим більше значення коефіцієнта детермінації і тим більше визначена регресія включеними в аналіз пояснювальними змінними.

ЗАВДАННЯ ДО ЛАБОРАТОРНОЇ РОБОТИ

Завдання 1. До завдання 1 та 2 лабораторної роботи 4 виконати такі пункти:

1. Використовуйте символи « * » та « : » для підгонки моделей лінійної регресії з ефектами взаємодії. Чи є якісь взаємодії статистично значимими? Чи змінився коефіцієнт детермінації R^2?

2. Спробуйте кілька різних перетворень змінних, наприклад, $\log x$, $\sqrt{x}$, x^2, $\mathrm{poly}(x, 5)$, $\log y$. Прокоментуйте свої висновки.

КОНТРОЛЬНІ ПИТАННЯ

1. Які моделі є нелінійними?
2. Які є класи нелінійних моделей?
3. Які перетворення предикторів можна використовувати для лінеаризації моделей? На що вони впливають?
4. Як оцінити достовірність нелінійної моделі?
5. Які значення може приймати кореляційне відношення?
6. Як додати ефекти взаємодії до моделі лінійної регресії в R?
7. Як перевірити вплив перетворення змінних на лінійність у регресійній моделі?
8. Як додати ефекти взаємодії до моделі лінійної регресії в R?

ЛАБОРАТОРНА РОБОТА № 6

Тема: якісні предиктори.

Мета: навчитися досліджувати якісні предиктори, існування зв'язку між якісними предикторами, будувати регресивні моделі з якісними предикторами та визначати за ними прогноз.

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Дотепер вважали, що всі змінні в нашій лінійній регресійній моделі є кількісними. Однак на практиці це не завжди так – часто деякі предиктори є якісними. Наприклад, набір даних Credit містить дані середнього боргу за кредитною картою (balance) для низки клієнтів банків, і так само кілька кількісних предикторов: age (вік), cards (кількість карт), education (кількість років, витрачених на освіту), income (дохід, тис. доларів), limit (кредитний ліміт) і rating (кредитний рейтинг).

Крім цих кількісних змінних, у нас є також чотири якісні змінні: gender (стать), student (чи клієнт студентів), married (чи клієнт в шлюбі) і ethnicity (європейського походження, афроамериканець або азіат).

Предиктори, що мають тільки два рівні

Уявімо, що ми хочемо з'ясувати відмінності за балансом на кредитній карті між чоловіками (Male) та жінками (Female), ігноруючи поки інші змінні. Якщо якісний предиктор (відомий також як «фактор») має тільки два рівні, або можливі значення, то додати його в регресійну модель дуже просто. До моделі вводиться індикаторна, або фіктивна, змінна, яка приймає тільки два можливі кількісні значення. Наприклад, на основі змінної gender (стать) ми можемо створити нову змінну виду:

$$x_i = \begin{cases} 1, \text{якщо } i \text{ (клієнт)} - \text{жінка;} \\ 0, \text{якщо } i \text{ (клієнт)} - \text{чоловік.} \end{cases}$$

і використовувати цю змінну як предиктор у рівнянні регресії. Це приводить до такої моделі:

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{якщо } i \text{ (клієнт)} - \text{жінка;} \\ \beta_0 + \epsilon_i, \text{якщо } i \text{ (клієнт)} - \text{чоловік.} \end{cases}$$

Тут β_0 можна інтерпретувати як середній баланс на кредитній карті у чоловіків, $\beta_0 + \beta_1$ – як середній баланс на кредитній карті у жінок, а β_1 – як середню різницю за балансом на кредитній карті між жінками і чоловіками.

У табл. 6.1 наведені оцінки коефіцієнтів та інша інформація за моделлю.

Таблиця 6.1 – Коефіцієнти регресії balance no gender, розраховані за методом найменших квадратів на основі даних Credit

	Коєфіцієнт	**Похибка**	**t**	**p**
Вільний член	509.80	33.13	15.389	<0.0001
Gerder (Female)	19.73	46.05	0.429	0.6690

Так, середній борг на кредитній карті у чоловіків оцінений у \$509,80, тоді як у жінок він виявився на \$19,73 вище, тобто \$509,80 + \$19,73 = \$529,53. Зауважте, однак, що p-значення у індикаторної змінної дуже високе. Це вказує на відсутність статистично значущої різниці між чоловіками і жінками за середнім балансом на кредитній карті.

Інші критерії визначення кореляції між якісними предикторами – рангова кореляція.

Коефіцієнт рангової кореляції Спірмена

У багатьох випадках результати спостережень подаються не у вигляді кількісних вимірювань, а у вигляді бальних оцінок (рангів). Наприклад, студенти у групі можуть бути впорядковані за номерами за середнім балом на сесії, країни – за кількістю населення, учасники конкурсу – за зайнятим місцем тощо. Інколи виникає можливість упорядкувати об'єкти дослідження за двома або більше показниками. Виникає задача дослідження кореляції цих показників.

Нехай n об'єктів дослідження, розташованих за рівнем якості, характеризуються парами рангів (x_i, y_i), i=1,2,…,*n*. Потрібно з'ясувати рівень кореляції між двома ознаками, x та y. Для цього використовують *коефіцієнт рангової кореляції Спірмена*.

Цей показник розраховують за формулою:

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)},$$

де $d = (x_i - y_i)$ – різниця рангів для і-го об'єкта спостереження.

Це значення можна використати як наближення коефіцієнта кореляції, він є менш точним, порівняно зі звичайним коефіцієнтом кореляції. У розрахунку не враховуються кількісні значення характеристик об'єктів, а лише їх порядок. Як і для звичайного коефіцієнта кореляції, значення R змінюється від -1 до 1. Чим ближче абсолютне значення коефіцієнта рангової кореляції до одиниці, тим більш щільним є зв'язок між факторами.

Приклад

У таблиці наведено дані про місця, що займають 8 провідних компаній галузі за собівартістю продукції (фактор x) та часткою ринку (фактор y). Обчислити коефіцієнт рангової кореляції Спірмена.

Таблиця 6.2 – Дані про розподіл компаній галузі за собівартістю продукції та часткою ринку

Підприємство	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Фактор x	8	3	1	4	2	7	5	6
Фактор y	3	5	6	7	8	4	1	2
$\boldsymbol{d = x - y}$	5	–2	–5	–3	–6	3	4	6

Маємо:

$$\sum_{i=1}^{8} d_i^2 = 25 + 4 + 25 + 9 + 36 + 9 + 36 + 9 + 16 + 16 = 140.$$

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} = 1\frac{6 \cdot 140}{8 \cdot 83} = -0{,}67.$$

Отримане значення коефіцієнта рангової кореляції Спірмена свідчить про наявність взаємозв'язку між собівартістю продукції компанії та часткою ринку. Під час цього спостерігається обернена залежність: зі зростанням собівартості продукції компанії її частка ринку зменшується.

Розглянемо методику оцінки зв'язку між якісними ознаками з використанням **коефіцієнтів спряження.**

В основі обчислення щільності зв'язку між атрибутивними (якісними) ознаками знаходиться побудова таблиці взаємного спряження (взаємозалежності) (табл. 6.3), у якій наведено комбінаційні розподіли сукупностей за факторною та результативною ознаками.

Таблиця 6.3 – Загальний вигляд таблиці взаємного спряження

Групи за ознакою *x*	**Групи за ознакою *y***						
	Група 1	**Група 2**	**…**	**Група *j***	**…**	**Група m_2**	**Разом**
Група 1	f_{11}	f_{12}	…	f_{1j}	…	f_{1m_2}	f_{10}
Група 2	f_{21}	f_{22}	…	f_{2j}	…	f_{2m_2}	f_{20}
…	…	…	…	…	…	…	…
Група i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{im_2}	f_{i0}
…	…	…	…	…	…	…	…
Група m_1	$f_{m_1 1}$	$f_{m_1 2}$	…	$f_{m_1 j}$	…	$f_{m_1 m_2}$	$f_{m_1 0}$
Разом	f_{01}	f_{11}	…	f_{0j}	…	f_{0m_2}	n

Величина f_{ij} – це число спостережень на перетині і-го рядка та j-го стовпця, тобто частота групи і у групі j, а f_{i0} та f_{0j} – відповідно підсумкові частоти за ознакою x та ознакою y. У випадку відсутності стохастичної залежності між ознаками частки умовних розподілів збігаються і дорівнюють часткам безумовного розподілу (часткам розподілу за підсумковим рядком). Розбіжність між фактичною кількістю спостережень у клітинках табл. 6.3 і теоретично можливою за повної відсутності зв'язку оцінюють за допомогою показника χ^2, який розраховують за формулою:

$$\chi^2 = n\left[\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{f_{ij}^2}{f_{i0}f_{j0}} - 1\right].$$

За відсутності зв'язку між ознаками $\chi^2 = 0$.

Для вимірювання щільності зв'язку між ознаками використовують кілька коефіцієнтів спряження.

Найчастіше використовують ***коефіцієнт Чупрова.*** Його обчислюють за формулою:

$$K_ч = \sqrt{\frac{\chi^2}{n\sqrt{(m_1-1)(m_2-1)}}}.$$

Тут n – кількість спостережень.

Якщо кількості виділених груп за кожною ознакою рівні, тобто $m_1 = m_2$ і між ознаками існує функціональний зв'язок, то коефіцієнт Чупрова дорівнює 1.

Проте, якщо $m_1 \neq m_2$, то значення коефіцієнта Чупрова відмінне від 1 навіть за наявності функціонального зв'язку між ознаками.

Модифікацією коефіцієнта Чупрова є ***коефіцієнт Крамера:***

$$K_K = \sqrt{\frac{\chi^2}{n(m-1)}}.$$

Тут $m = min(m_1, m_2)$.

Оцінити щільність зв'язку між якісними ознаками можна також за допомогою ***коефіцієнта Пірсона:***

$$K_П = \sqrt{\frac{\chi^2}{n+\chi^2}}.$$

Значення коефіцієнтів Чупрова, Крамера та Пірсона коливаються у межах від 0 до 1. Коефіцієнт Чупрова враховує кількість виділених груп за кожною ознакою і дає найбільш обережну оцінку щільності зв'язку. Якщо значення цього коефіцієнта $К_ч = 0{,}3$, то можна говорити про помірний або щільний зв'язок між ознаками. Перевірка істотності зв'язку здійснюється на основі χ^2-критерію з $V = (m_1 - 1)(m_2 - 1)$ ступенями вільності.

ЗАВДАННЯ ДО ЛАБОРАТОРНОЇ РОБОТИ

Завдання 1. У цьому завданні використовуються дані Carseats (можна скачати за посиланням https://www.kaggle.com/datasets/huhao05133/carseats). У завданні потрібно передбачити змінну Sales (продажі дитячих автокрісел) у 400 регіонах на основі низки предикторів.

Завдання 2. Знайдіть модель множинної лінійної регресії, яка включає всі змінні набору даних. Прокоментуйте отримані результати, звернувши увагу на індикаторні змінні.

КОНТРОЛЬНІ ПИТАННЯ

1. Які з предикторів є кількісними?
2. Скільки та які значення приймають якісні предиктори?
3. Які перетворення можна виконувати з якісними предикторами?
4. Як оцінити щільність зв'язку між якісними предикторами?
5. Які значення приймають коефіцієнти Спірмена та Кендела?
6. Що є коефіцієнтом спряження?
7. Як вводяться у модель предиктори, що мають тільки два значення?
8. Як вводяться у модель предиктори, що мають три і більше значень?
9. За допомогою функції contrasts() показати, як саме R кодує індикаторні змінні.
10. Як визначити якісні предиктори в R?
11. Що таке dummy variables (фіктивні змінні) та як їх використовувати в R?

ІНДИВІДУАЛЬНІ ВАРІАНТИ ДОСЛІДЖУВАНОЇ (ОРГАНІЗАЦІЙНОЇ, СОЦІАЛЬНО-ЕКОНОМІЧНОЇ) СИСТЕМИ

1. Будівельна компанія.
2. Житлово-комунальне підприємство.
3. Автотранспортне підприємство.
4. Фірма з організації вантажоперевезень.
5. Виробниче підприємство.
6. Заклад вищої освіти (академія).
7. Школа.
8. Велика міжнародна корпорація.
9. Готельний комплекс.
10. Фірма-розробник програмного забезпечення.
11. Лікарняний комплекс.
12. Торговельне підприємство.
13. Телерадіокомпанія.
14. Спортивно-оздоровчий комплекс.
15. Аграрне підприємство.
16. Поліграфічний комбінат.
17. Нафтопереробне підприємство.
18. Газотранспортний консорціум.
19. Логістичний центр.
20. Сервісний центр обслуговування.
21. Консалтингова компанія.
22. ІТ-компанія.
23. Розважальний центр.
24. Видавництво.
25. Інформаційне агентство.
26. Магазин комп’ютерної техніки.
27. Магазин меблів.
28. Квітковий салон.
29. Косметичний салон.
30. Магазин одягу.
31. Спорттовари.
32. Ресторан.
33. Банк.
34. Аптека.
35. Магазин іграшок.
36. Магазин побутової техніки.
37. Магазин мобільних телефонів.
38. Магазин будівельних товарів.
39. Продуктовий магазин.
40. Книжковий магазин.
41. Туристична агенція.
42. Склад товарів.

43. Рекламна компанія.
44. Медичний центр.
45. Ювелірний магазин.
46. Агентство нерухомості.
47. Автосалон.
48. Весільний салон.
49. Агрофірма.
50. Автовокзал.

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. The R Project for Statistical Computing. URL: https://www.r-project.org/

2. An introduction to statistical learning (with applications in R) / G. James, D. Witten, T. Hastie, & R. Tibshirani. Springer New York, NY. 2021. 607 p.

3. Методичні вказівки до виконання практичних робіт з дисципліни «Інтелектуальний аналіз даних» для студентів спеціальності 122 Комп'ютерні науки / уклад. Т. В. Федорончак. Запоріжжя: ЗНТУ, 2017. 44 с.

4. Моделі та методи соціально-економічного прогнозування: підручник / В. М. Геєць, Т. С. Клебанова, О. І. Черняк та ін. Харків: ВД «ІНЖЕК», 2005. 396 с.

5. Єріна А. М., Єрін Д. Л. Статистичне моделювання та прогнозування: підручник. Київ: КНЕУ, 2014. 348 с.

ДЛЯ НОТАТОК

ДЛЯ НОТАТОК

Навчальне видання

Волонтир Людмила Олексіївна
Потапова Надія Анатоліївна
Хмелівський Юрій Сергійович

СТАТИСТИЧНЕ НАВЧАННЯ. ЧАСТИНА 1

Методичні вказівки
для виконання лабораторних робіт
для здобувачів ОС «Бакалавр»
спеціальності 122 Комп’ютерні науки

Редактор О. А. Солдатова
Технічний редактор Т. О. Важеніна-Гопрак

Підписано до друку 16.12.2024.
Формат 60 × 84/16. Папір офсетний.
Друк – цифровий. Умовн. друк. арк. 2,79.
Тираж 30. Зам. 50.

Донецький національний університет імені Василя Стуса
21021, м. Вінниця, 600-річчя, 21.
Свідоцтво про внесення суб’єкта видавничої справи
до Державного реєстру
серія ДК № 5945 від 15.01.2018