Практична робота № 3

Тема: Простий регресійний аналіз

Мета роботи: Знайти коефіцієнти поліноміальної та нелінійних регресійних залежностей для отриманих унаслідок експерименту масивів даних *X*, *Y*, де *Y* залежить від *X*. Перевірити адекватність моделі критерієм Фішера.

Стислі теоретичні відомості

Лінійна регресійна модель

Прості парні регресійні моделі встановлюють лінійну залежність між двома змінними, наприклад, витратами на відпустку та складом родини; витратами на рекламу та обсягом продукції, що виробляється. При цьому одна із змінних вважається залежною змінною (y) та розглядається як функція від незалежної змінної (x). Аналітична залежність $y = f(x)$ має назву регресійної залежності, а Її пошук – простого регресійного аналізу.

У загальному вигляді проста лінійна регресійна модель записується таким чином: $y = b_0 + b_1 x + e$,

де y – вектор спостережень за залежною змінною $y = \{y_1, y_2,\ldots,y_n\}$; x – вектор спостережень за незалежною змінною $x = \{x_1, x_2,\ldots,x_n\}$; b_0, b_1 – невідомі параметри регресійної моделі; e – вектор випадкових величин (помилок) $e = \{e_1, e_2,\ldots,e_n\}$.

Оцінювання параметрів лінійної регресії за допомогою методу найменших квадратів

Щоб мати явний вигляд залежності, необхідно знайти (оцінити) невідомі параметри b_0, b_1 цієї моделі. Розглянемо лінійну залежність $\hat{y} = b_0 + b_1 x$ як найбільш поширену. Нижче буде вказано зведення нелінійних регресійних залежностей до лінійних. Найбільш поширеним критерієм знаходження коефіцієнтів регресійної залежності є критерій мінімізації суми квадратів відхилень значень регресійної залежності $\hat{y}_i$ у точках спостереження x_i від результатів спостереження y_i.

Відхилення (помилки) теоретичних значень від практичних можна виразити формулою

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i, i = \overline{1_i, n},$$

де $\hat{y}_i$ – i-та точка на прямій, яка відповідає значенню x_i (рис. 3).

Відхилення або помилки інколи називають залишками. Логічно, що треба проводити пряму таким чином, щоб сума квадратів помилок була мінімальною. У цьому й полягає **критерій найменших квадратів:** невідомі параметри b_0 та b_1 знаходяться таким чином, щоб мінімізувати

$$\sum_{i=1}^{n} e_i^2 .$$

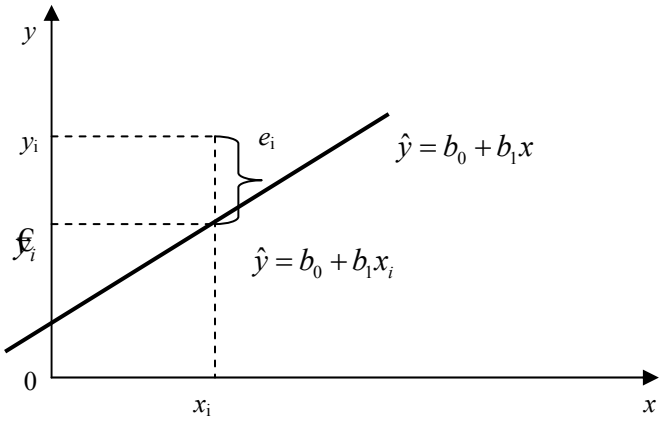

Рис. 3. Відхилення теоретичних значень від фактичних

Справді, за ним ми маємо:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = f(b_0, b_1) \to \min.$$

Мінімізація цієї функції, яка є функцією від двох невідомих b_0, b_1, дає таку формулу для параметра b_1 (нахилу):

$$b_1 = \frac{\frac{1}{n}(\sum_{i=1}^{n} x_i y_i) - \bar{x} \cdot \bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2},$$

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

При цьому параметр b_0 (перетин) знаходиться за формулою:

$$b_0 = \bar{y} - b_1 \bar{x}.$$

Оцінювання параметрів нелінійної регресії за допомогою методу найменших квадратів

Припустимо, що внаслідок експерименту ми отримали два масиви даних *X, Y,* де *Y* залежить від *X*. Аналітична залежність $y = f(x)$ має назву регресійної залежності. Частіше всього використовують такі регресійні залежності:

$y = b_0 + b_1 x + b_2 x^2$ – поліноміальна залежність;

$y = b_0 + b_1 x$ – лінійна залежність,

$y = b_0 x^{b_1}$ – ступенева залежність,

$y = b_0 b_1^x$ – показникова залежність,

$y = e^{b_0 * b_1^x}$ – крива Гомперця,

$y = \frac{1}{b_0 + b_1 x}$ – зворотна залежність.

Визначення регресійних залежностей зводиться до визначення їх коефіцієнтів таким чином, щоб залежність максимально задовольняла усім даним експерименту. Критерієм такої задоволеності є сума квадратів відстаней точок експерименту від теоретичної залежності (відстані беруться у квадраті, щоб від'ємні і додатні значення не компенсували один одного).

Розглянемо пошук коефіцієнтів поліноміальної регресійної залежності методом найменших квадратів (МНК).

Уводимо два масиви даних експерименту X, Y, де Y залежить від X.

$$X := \begin{pmatrix} 0{,}25 \\ 0{,}5 \\ 0{,}75 \\ 1 \\ 1{,}25 \\ 1{,}5 \\ 1{,}75 \\ 2 \\ 2{,}25 \\ 2{,}5 \end{pmatrix}, Y := \begin{pmatrix} 0{,}4 \\ 0{,}5 \\ 0{,}9 \\ 1{,}28 \\ 1{,}6 \\ 1{,}66 \\ 2{,}02 \\ 2{,}4 \\ 2{,}61 \\ 2{,}81 \end{pmatrix}$$

Поліноміальну регресійну залежність шукаємо у вигляді:

$$y(x) := b_0 + b_1 \cdot x + b_2 \cdot x^2.$$

Визначаємо функцію, яка є сумою квадратів відхилень результатів експерименту від тих даних, які ми отримуємо завдяки моделі для кожної точки:

$$f(b_0, b_1, b_2) := \sum_{i=0}^{9} (Y_i - (b_0 + b_1 X_i + b_2 X_i^2))^2,$$

де X_i, Y_i – i-й елемент масивів X, Y.

Шукаємо точку з координатами b_0, b_1, b_2, яка визначає точку мінімуму функції $f(b_0, b_1, b_2)$. Відомо, що в точці мінімуму частинні похідні функції по кожному з аргументів дорівнюють нулю.

Внаслідок того, що:

$$\frac{\partial}{\partial a_0} f(b_0, b_1, b_2) = \sum_{i=0}^{9} 2(Y_i - (b_0 + b_1 X_i + b_2 X_i^2))(-1),$$

$$\frac{\partial}{\partial a_1} f(b_0, b_1, b_2) = \sum_{i=0}^{9} 2(Y_i - (b_0 + b_1 X_i + b_2 X_i^2))(-X_i),$$

$$\frac{\partial}{\partial a_2} f(b_0, b_1, b_2) = \sum_{i=0}^{9} 2(Y_i - (b_0 + b_1 X_i + b_2 X_i^2))(-X_i^2),$$

система алгебраїчних рівнянь для знаходження коефіцієнтів має вигляд:

$$\sum_{i=0}^{9} 2(Y_i - (b_0 + b_1 X_i + b_2 X_i^2))(-1) = 0;$$

$$\sum_{i=0}^{9} 2(Y_i - (b_0 + b_1 X_i + b_2 X_i^2))(-X_i) = 0;$$

$$\sum_{i=0}^{9} 2(Y_i - (b_0 + b_1 X_i + b_2 X_i^2))(-X_i^2) = 0.$$

Якщо ми маємо справу з експоненціальною, ступеневою, та показниковою залежностями, то шляхом логарифмування зводимо їх до лінійної та використовуємо формули, які наведені вище. Приклади пошуку коефіцієнтів таких залежностей наведені далі. Після знаходження коефіцієнтів регресійної залежності проводять аналіз дисперсій та перевірку адекватності моделі.

Простий ANOVA-аналіз регресії: аналіз дисперсій

Таблиця дисперсійного аналізу (*ANOVA)* має такий вигляд (табл. 2) [3]:

Таблиця 2

Джерело варіації	**Кількість ступенів вільності**	**Суми квадратів**	**Середні квадрати**
Зумовлене регресією	$k-1$	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	$MRS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \Big/ k-1$
Не пояснюване за допомогою регресії	$n-k$	$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2$	$MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ n-k$
Загальне, скориговане на середнє	$n-1$	$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$	

Можна довести, що: *SST* = *SSR* + *SSE*, отже

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST} = R^2 + \frac{SSE}{SST}$$

де $R^2 = \frac{SSR}{SST}$ – коефіцієнт детермінації моделі $R^2 = 1 - \frac{SSE}{SST}$.

Знаходять коефіцієнт детермінації моделі**,** який повинен як можна більше наближатися до 1.

Перевірка простої регресійної моделі на адекватність за *F*-критерієм Фішера

Найбільш поширеним із критеріїв перевірки адекватності моделі є критерій Фішера. Перевірка моделі на адекватність за F-критерієм Фішера складається з певних етапів:

1. На першому етапі розраховуємо величину, так зване F-відношення:

$$F_{(k-1,n-k)} = \frac{MSR}{MSE} = \frac{\frac{1}{(k-1)} * \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\frac{1}{n-k} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} ,$$

де *MSR* – середній квадрат, який можна пояснити з регресійної моделі; *MSE* – середній квадрат помилок; $k - 1$, $(n - k)$ – ступені вільності, відповідно пов'язані з *MSR* і *MSE*, k – кількість коефіцієнтів у регресійній моделі.

2. На другому етапі задаємо рівень значимості α або α*100%. Наприклад, якщо ми вважаємо, що можлива помилка α для нас становить 0.05 (або 5%), це означає, що ми можемо помилитися не більш ніж у 5% випадків, а в 95% (або в 100*(1 – α)%) наші висновки будуть правильними.

3. На третьому етапі за статистичними таблицями *F*-розподілу Фішера з $k - 1$, $(n - k)$ ступенями вільності та рівним значимості 100*(1 – α)% знаходимо критичне значення ($F_{кр}$), яке дорівнює квантилю (1 – α) розподілу Фішера з ($k - 1$, $n - k$) ступенями вільності, тобто такому значенню випадкової величини, яке є коренем інтегрального рівняння

$$\int_0^{F_{кр}} f(x)dx = (1-\alpha)$$

та відповідає межі інтервалу, для якого ймовірність влучення випадкової величини підпорядкована закону розподілу Фішера (1 – α).

4. Якщо розраховане нами значення $F > F_{кр}$, то ми можемо довіряти моделі більш ніж на (1 – а), тому відкидаємо гіпотезу H_0, що $\beta_1 = 0$ (або що $(\hat{y}_i = \overline{y})$, з ризиком помилитися не більше ніж 5%.

Отже, якщо $F > F_{кр}$, то побудована нами регресійна модель адекватна реальній дійсності.

Після перевірки адекватності моделі у цілому перевіряють на значимість кожен з коефіцієнтів моделі за допомогою *t*-тесту. Як це робиться, буде показано далі на прикладі множинної регресійної моделі, де це більш необхідно.

Приклади виконання у пакеті Mathcad

1. *Знаходження коефіцієнтів лінійної та поліноміальної регресійної залежності*

Уводимо два масиви даних експерименту *X*, *Y*, де *Y* залежить від *X*.

$$X := \begin{pmatrix} 0{,}25 \\ 0{,}5 \\ 0{,}75 \\ 1 \\ 1{,}25 \\ 1{,}5 \\ 1{,}75 \\ 2 \\ 2{,}25 \\ 2{,}5 \end{pmatrix}, Y := \begin{pmatrix} 0{,}4 \\ 0{,}5 \\ 0{,}9 \\ 1{,}28 \\ 1{,}6 \\ 1{,}66 \\ 2{,}02 \\ 2{,}4 \\ 2{,}61 \\ 2{,}81 \end{pmatrix}$$

Поліноміальну регресійну залежність шукаємо у вигляді:

$$y(x) = b_0 + b_1 \times x + b_2 \cdot x^2.$$

Пошук коефіцієнтів поліноміальної регресійної залежності у середовищі пакета Mathcab можливо також виконувати за допомогою функції **regress(*X*, *Y*, 2).**

Аргументами функції є вектор незалежних даних *X*, вектор залежних даних *Y*, ступінь полінома 2. Виходом функції є вектор, третій елемент якого є ступенем полінома, останні елементи – коефіцієнти b_0, b_1, b_2.

Приклад виконання завдання у середовищі пакета Mathcad:

***Z*:= regress(*X*, *Y*, 2).**

$$Z = \begin{bmatrix} 3 \\ 3 \\ 2 \\ 0.163 \\ 1.64 \\ 1.662 \end{bmatrix}$$

Тобто $y(x) = 0.163 + 1.664 \cdot x + 1.662 \cdot x^2$.

Лінійну регресійну залежність $y(x) = b_0 + b_1 x$ у середовищі пакета Mathcad можливо знайти за допомогою функції regress (***Z*:= regress(*X*, *Y*, 1)).** Її також можливо знайти за допомогою функцій **intercept, slope.**

b_0:= intercept(X,Y) b_1:= slope(X,Y).

2. *Знаходження коефіцієнтів ступеневої регресійної залежності*

Уводимо 2 масиви даних експерименту *X*, *Y*, де *Y* залежить від *X*.

$$X := \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{pmatrix}, Y := \begin{pmatrix} 2 \\ 2,8 \\ 3 \\ 4 \\ 4,5 \\ 5 \\ 5,3 \\ 5,6 \\ 6 \\ 5 \end{pmatrix}$$

Нелінійну регресійну залежність шукаємо у вигляді $y = b_0 x^{b1}$.

При логарифмуванні регресійної залежності маємо:

$$ln\ y(x) = ln\ b_0 + b_1\ ln\ (x).$$

Позначимо: $y1 = ln(y)$; $b_{01} = ln\ b_0$; $x1 = ln(x)$.

Тоді $X1_i = ln(X_i)$; $Y1_i = ln(Y_i)$; $Y1_i(X1_i) := b_{01} + b_1 \cdot X1_i$.

Знаходження коефіцієнтів регресійної залежності та перевірка адекватності моделі критерієм Фішера наведені на рис. 4.

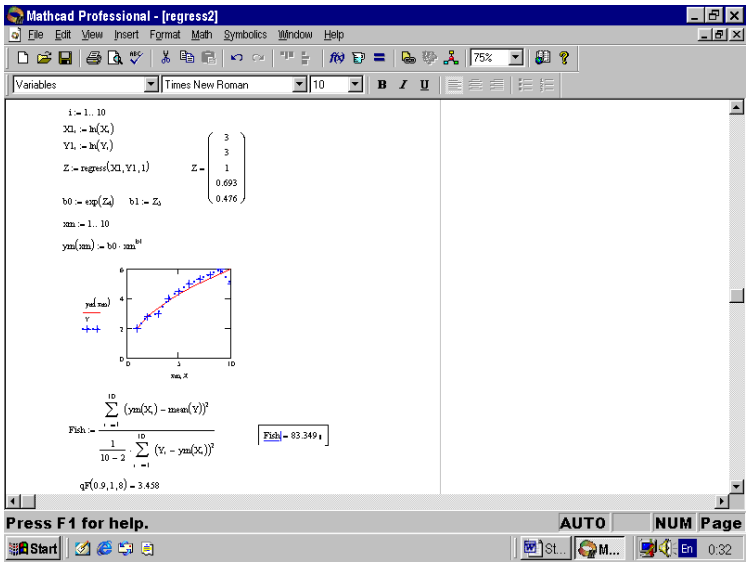

Рис. 4

3. *Знаходження коефіцієнтів експоненціально-показникової регресійної залежності*

$$y = e^{b_0^* (b_1)^x}.$$

Уводимо 2 масиви даних експерименту X, Y, де Y залежить від X.

$$X := \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{pmatrix}, Y := \begin{pmatrix} 1{,}5 \\ 1{,}7 \\ 2 \\ 2{,}2 \\ 2{,}8 \\ 3{,}7 \\ 7 \\ 10 \\ 13 \\ 24 \end{pmatrix}$$

При подвійному логарифмуванні маємо:

$$ln(ln(y)) = ln\ b_0 + x\ ln(b_1).$$

Знаходження коефіцієнтів регресійної залежності та перевірка адекватності моделі критерієм Фішера наведені на рис. 5. Як вказано на рис. 5, критерій Фішера (Fish) значно більше квантиля Фішера, який відповідає ймовірності 0.95 (рівень довіри моделі) та ступеням вільності 1 і 8.

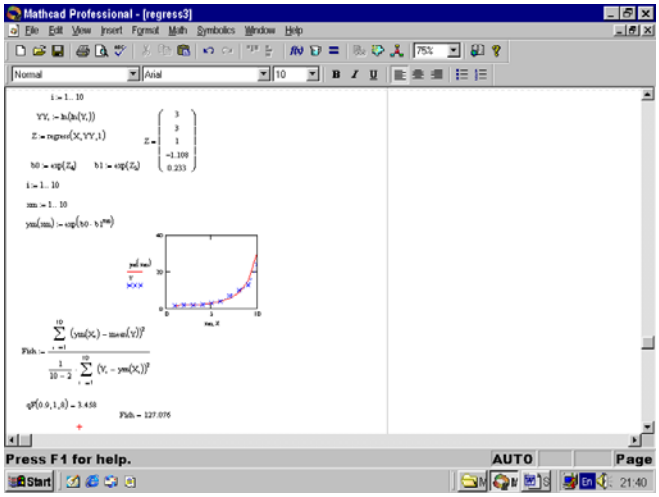

Рис. 5

Контрольні запитання

1. Яка основна ідея пошуку коефіцієнтів регресійної залежності методом найменших квадратів?
2. Який вигляд має система алгебраїчних рівнянь для пошуку коефіцієнтів поліноміальної регресійної залежності методом найменших квадратів?
3. Який вигляд має система алгебраїчних рівнянь для пошуку коефіцієнтів лінійної регресійної залежності методом найменших квадратів?
4. За допомогою яких функцій пакета Mathcad знаходяться коефіцієнти лінійної регресійної залежності?
5. Як знаходяться коефіцієнти ступеневої регресійної залежності?
6. Як знаходяться коефіцієнти показникової регресійної залежності?
7. Як знаходяться коефіцієнти експоненціально-показникової регресійної залежності?