

Лекція 8. Наближення функцій за табличними значеннями.

ПЛАН

1. Постановка задачі та вибір типу згладжувальної кривої
2. Суть методу найменших квадратів
3. Лінеаризація та відшукування параметрів лінійної функції
4. Кореляційний та регресійний аналіз.

8.1. Постановка задачі та вибір типу згладжувальної кривої

Нехай на підставі експерименту отримано n значень функції: y_1, y_2, \dots, y_n при відповідних значеннях аргументу x_1, x_2, \dots, x_n . Потрібно за даними значеннями змінних y та x встановити функціональну залежність між ними, тобто в рівності $y = \varphi(x)$ (8.1)

визначити φ . При цьому функція φ має "найкращим чином" описувати відповідний процес.

Наближений вигляд згладжувальної функції φ можна підібрати, виходячи з теоретичних міркувань або ж за характером розташування на координатній площині експериментальних точок $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$.

Так, якщо, наприклад, експериментальні точки розташовані на площині xOy так, як на рис.8, то, враховуючи похибки експерименту, природно допустити, що в якості функції φ можна взяти наближено лінійну функцію

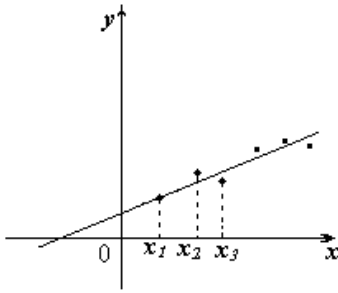
$$y = ax + b. \quad (8.2)$$

Якщо ж точки $(x_i; y_i)$, $i = 1, 2, \dots, n$ розташовані так, як на рис.2, то згладжувальною функцією може слугувати гіперболічна залежність

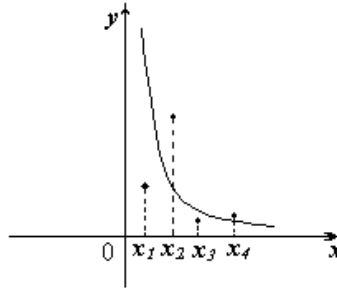
$$y = \frac{a}{x} + b \quad (8.3)$$

або

$$y = \frac{1}{ax + b}. \quad (8.4)$$



мал.1



мал.2

Залежності (2) – (4) визначаються двома невідомими параметрами, тобто є двохпараметричними, типу

$$y = \varphi(x, a, b). \quad (8.5)$$

Згладжувальна функція може визначатися трьома

$$y = \varphi(x, a, b, c) \quad (8.6)$$

та більше параметрами; зокрема, відома квадратична залежність

$$y = ax^2 + bx + c \quad (8.7)$$

є конкретизацією (8.6).

8.2. Суть методу найменших квадратів

При вибраному вигляді згладжувальної функції (8.5) потрібно підібрати параметри a та b так, щоб сума квадратів відхилень y_i від $\varphi(x_i, a, b)$ (значень функції (8.5) в точках $x_i, i = 1, 2, \dots, n$) була найменшою, тобто

$$S = S(a, b) = \sum_{i=1}^n [y_i - \varphi(x_i, a, b)]^2 \rightarrow \min_{a, b} \quad (8.8)$$

(В цьому розумінні функція φ за методом найменших квадратів "найкращим чином" описує відповідний процес).

Задача (8.8) – це задача на безумовний екстремум функції двох змінних $S = S(a, b)$. На підставі необхідної умови екстремуму параметри a та b знаходяться з умови

$$\begin{cases} \frac{\partial S}{\partial a} = 0, \\ \frac{\partial S}{\partial b} = 0. \end{cases} \quad (8.9)$$

Якщо в якості згладжувальної функції вибрано (8.6), то за методом найменших квадратів параметри a, b, c підбираються за критерієм

$$S = S(a, b, c) = \sum_{i=1}^n [y_i - \varphi(x_i, a, b, c)]^2 \rightarrow \min_{a, b, c}, \quad (8.10)$$

тобто є розв'язками такої системи рівнянь:

$$\begin{cases} \frac{\partial S}{\partial a} = 0, \\ \frac{\partial S}{\partial b} = 0, \\ \frac{\partial S}{\partial c} = 0. \end{cases} \quad (8.11)$$

Надалі співвідношення (8.8) - (8.11) конкретизуються для деяких певних типів згладжувальних функцій (8.5), (8.6).

8.3. Лінеаризація та відшукування параметрів лінійної функції

Для лінійної функції (8.3)

$$S(a, b) = \sum_{i=1}^n [y_i - (ax_i + b)]^2,$$

система (8.9) набуває вигляду

$$\begin{cases} a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i, \\ a \sum_{i=1}^n x_i + b \cdot n = \sum_{i=1}^n y_i. \end{cases} \quad (8.12)$$

(8.12) називається *нормальною системою методу найменших квадратів для відшукування параметрів лінійної залежності* (8.3).

Це звичайна система двох лінійних рівнянь з двома невідомими a та b .

Лінеаризація та відшукування параметрів нелінійних згладжувальних двохпараметричних функцій

У випадку нелінійної залежності між змінними x та y можна застосувати так званий *метод лінеаризації*. Суть цього методу полягає в тому, що вводяться нові змінні ξ, η , а, можливо, і параметри, α, β , відносно яких залежність стає вже лінійною:

$$\eta = \alpha \xi + \beta. \quad (8.13)$$

Отже, щоб скласти рівняння нелінійної згладжувальної кривої, потрібно лінеаризувати цю нелінійну залежність, тобто звести її до вигляду (8.13). У відповідності із заміною змінних слід перерахувати вихідні експериментальні дані (кожній парі (x_i, y_i) поставити у відповідність нову пару (ξ_i, η_i) , $i=1,2,\dots,n$) та для лінійної залежності (8.13) скласти нормальну систему відносно α, β

$$\begin{cases} \alpha \sum_{i=1}^n \xi_i^2 + \beta \sum_{i=1}^n \xi_i = \sum_{i=1}^n \xi_i \eta_i, \\ \alpha \sum_{i=1}^n \xi_i + \beta \cdot n = \sum_{i=1}^n \eta_i. \end{cases} \quad (8.14)$$

Розв'язавши (8.14), потрібно повернутися до "старих" параметрів a, b та записати остаточне рівняння згладжувальної кривої типу (8.5).

Відшукання параметрів квадратичної функції

Для знаходження параметрів a, b, c квадратичної функції (8.7) слід скористатися критерієм

$$S = S(a, b, c) = \sum_{i=1}^n [y_i - (ax_i^2 + bx_i + c)]^2 \rightarrow \min_{a, b, c}, \quad (8.15)$$

тобто, за (11), розв'язати лінійну систему трьох лінійних рівнянь з трьома

невідомими a, b, c

$$\begin{cases} a \sum_{i=1}^n x_i^4 + b \sum_{i=1}^n x_i^3 + c \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i^2 y_i, \\ a \sum_{i=1}^n x_i^3 + b \sum_{i=1}^n x_i^2 + c \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i, \\ a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i + c \cdot n = \sum_{i=1}^n y_i. \end{cases} \quad (8.16)$$

8.4. Кореляційний та регресійний аналіз.

1. Залежність між випадковими величинами

У багатьох наукових дослідженнях виникає необхідність проводити одночасно спостереження над кількома випадковими величинами, щоб встановити та оцінити їх взаємозв'язок.

Дві випадкові величини X та Y можуть бути зв'язані або функціональною, або статистичною залежністю, або ж бути взагалі незалежними.

Якщо кожному можливому значенню x величини X відповідає певне значення y іншої величини Y , то говорять, що випадкові величини X та Y зв'язані *функціональною залежністю*.

Строга функціональна залежність реалізується дуже рідко, оскільки випадкові величини X та Y (або ж одна з них) зазнають впливу багатьох випадкових факторів, серед яких можуть бути і спільні.

В цьому випадку між двома величинами X та Y виникає *статистична залежність*.

2. Вибіркові рівняння регресії

Для випадку кореляційної залежності, якщо величина X прийняла значення x , то математичне сподівання величини Y є при цьому функцією від x :

$$M_x Y = f(x) \quad (8.17)$$

Рівняння (1) називається *рівнянням регресії Y на X* .

Оскільки математичне сподівання є істинним (справжнім) значенням величини Y , що спостерігається, то рівняння регресії (8.17) дає справжню залежність між величинами X та Y . Тому кінцевою метою багатьох досліджень є знаходження вибіркового рівняння регресії (8.17), яке прийнято записувати у вигляді $\bar{y}_x = f(x)$.

Тут \bar{y}_x - *умовне середнє* (це середнє арифметичне значень випадкової величини Y , що відповідають значенню $X = x$); f - *функція регресії Y на X* .

$$\text{Аналогічно рівняння } \bar{x}_y = \varphi(y) \quad (8.19)$$

називають вибірковим *рівнянням регресії X на Y* . При цьому рівнянням (8.18), (8.19) в загальному випадку відповідають дві різні лінії на площині XOY .

В математичній статистиці при дослідженні залежності однієї кількісної ознаки Y від іншої ознаки X виникають дві основні задачі:

1) знайти наближену функцію регресії, що характеризує основну тенденцію залежності Y від X (або X від Y) та належить одному з відомих

типів функцій (лінійна, квадратична, показникова і т.д.);

2) оцінити силу, тісноту цієї залежності.

3. *Відшукування параметрів вибіркового рівняння регресії за незгрупованими даними*

Нехай вивчається вибірка об'єму n з двох кількісних ознак X та Y : $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$, причому значення $x_i; y_i, i = \overline{1, n}$ зустрічається по одному разу.

В цьому випадку немає необхідності групувати дані та використовувати поняття умовної середньої. Тому шукане рівняння регресії можна записати так:

$$y = f(x) \quad (8.21)$$

або

$$x = \varphi(y). \quad (8.22)$$

Рівняння згладжувальної кривої і буде наближеним рівнянням регресії.

4. *Вибірковий коефіцієнт кореляції*

Найбільш поширеним на практиці є випадок лінійного зв'язку досліджуваних ознак X та Y . Тому важливою характеристикою вибірки є статистична оцінка для коефіцієнта кореляції генеральної сукупності, так званий *вибірковий коефіцієнт кореляції*:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (8.23)$$

де $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ - відповідні середні значення (вибіркові середні).

Вибірковий коефіцієнт кореляції, як оцінка коефіцієнта кореляції генеральної сукупності, слугує для вимірювання лінійного зв'язку між величинами - кількісними ознаками X та Y .

Властивості вибіркового коефіцієнта кореляції:

1. $-1 \leq r_{xy} \leq 1$.

2. Чим більша величина r_{xy} , тим тісніший зв'язок між досліджуваними ознаками X та Y .

3. Якщо $|r_{xy}| = 1$, то кореляційна залежність між X та Y стає лінійною функціональною.

4. Якщо $|r_{xy}| = 0$, то між досліджуваними ознаками X та Y немає лінійної кореляційної залежності, але умова $r_{xy} = 0$ не виключає існування будь-якої іншої кореляційної залежності (параболічної, показникової і т.д.). Якщо з деяких теоретичних міркувань заздалегідь відомо, що величини X та Y мають нормальний розподіл, то рівність $r_{xy} = 0$ свідчить про відсутність будь-якої залежності між ознаками X та Y (тобто величини X , Y - незалежні).

5. Рівняння лінійної регресії

У випадку лінійної згладжувальної функції

$$y = ax + b \quad (8.24)$$

параметри a, b , згідно з методом найменших квадратів, визначаються за формулами

$$a = \frac{\sigma_x}{\sigma_y} r_{xy}, \quad b = \bar{y} - a\bar{x} \quad (8.25)$$

де r_{xy} - вибірковий коефіцієнт кореляції (8.23);

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2}, \quad \sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2} \quad \text{- середні квадратичні відхилення.}$$

Кутовий коефіцієнт a прямої (8.24) називають вибірковим *коефіцієнтом регресії Y на X* .

Рівняння лінійної регресії Y на X прийнято записувати у вигляді

$$y - \bar{y} = r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x}). \quad (8.26)$$

Аналогічно можна знайти вибіркове рівняння прямої лінії регресії X на Y :

$$x - \bar{x} = r_{xy} \frac{\sigma_x}{\sigma_y} (y - \bar{y}). \quad (8.27)$$

Кореляційний та регресійний аналіз.

Залежність між випадковими величинами

У багатьох наукових дослідженнях виникає необхідність проводити одночасно спостереження над кількома випадковими величинами, щоб встановити та оцінити їх взаємозв'язок.

Дві випадкові величини X та Y можуть бути зв'язані або функціональною, або статистичною залежністю, або ж бути взагалі незалежними.

Якщо кожному можливому значенню x величини X відповідає певне значення y іншої величини Y , то говорять, що випадкові величини X та Y зв'язані *функціональною залежністю*.

Строга функціональна залежність реалізується дуже рідко, оскільки випадкові величини X та Y (або ж одна з них) зазнають впливу багатьох випадкових факторів, серед яких можуть бути і спільні.

В цьому випадку між двома величинами X та Y виникає *статистична залежність*.

Залежність між випадковими величинами X та Y , за якою кожному значенню однієї величини відповідає розподіл іншої, називається *статистичною*.

Зокрема, якщо кожному можливому значенню однієї величини ставиться у відповідність середнє значення іншої, то така статистична залежність називається *кореляційною*.

Вибіркові рівняння регресії

Для випадку кореляційної залежності, якщо величина X прийняла значення x , то математичне сподівання величини Y є при цьому функцією від x :

$$M_x Y = f(x) \quad (8.28)$$

Оскільки математичне сподівання є істинним (справжнім) значенням величини Y , що спостерігається, то рівняння регресії (8.28) дає справжню залежність між величинами X та Y . Тому кінцевою метою багатьох

досліджень є знаходження вибіркового рівняння регресії (8.28), яке прийнято записувати у вигляді

$$\bar{y}_x = f(x). \quad (8.29)$$

Тут \bar{y}_x - умовне середнє (це середнє арифметичне значень випадкової величини Y , що відповідають значенню $X = x$); f - функція регресії Y на X .

Графік (8.29) називають *вибірковою лінією регресії Y на X* .

Аналогічно рівняння

$$\bar{x}_y = \varphi(y) \quad (8.30)$$

називають вибірковим *рівнянням регресії X на Y* . При цьому рівнянням (10), (11) в загальному випадку відповідають дві різні лінії на площині XOY .

В математичній статистиці при дослідженні залежності однієї кількісної ознаки Y від іншої ознаки X виникають дві основні задачі:

1) знайти наближену функцію регресії, що характеризує основну тенденцію залежності Y від X (або X від Y) та належить одному з відомих типів функцій (лінійна, квадратична, показникова і т.д.);

2) оцінити силу, тісноту цієї залежності.

Відшукування параметрів вибіркового рівняння регресії за незгрупованими даними

Нехай вивчається вибірка об'єму n з двох кількісних ознак X та Y : $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$, причому значення $x_i; y_i, i = \overline{1, n}$ зустрічається по одному разу.

В цьому випадку немає необхідності групувати дані та використовувати поняття умовної середньої. Тому шукане рівняння регресії можна записати т

$$y = f(x) \quad (8.31)$$

або

$$x = \varphi(y). \quad (8.32)$$

Наближений вигляд згладжувальної функції (f або φ) можна підібрати, зокрема, за характером розташування на координатній площині XOY точок $x_i; y_i, i = \overline{1, n}$. При вибраному вигляді згладжувальної функції $y = f(x, a, b)$ ($y =$

$f(x,a,b,c)$ невідомі параметри $a, b(a,b,c)$ знаходять за методом найменших квадратів, розв'язуючи відповідну систему лінійних алгебраїчних рівнянь.

Рівняння згладжувальної кривої і буде наближеним рівнянням регресії.