Тема 4

Метод найменших квадратів

Лектор: **к.е.н., доц., доцент кафедри вищої математики, економетрії і статистики ДЕМЧИШИН М.Я.**

План

4.1. Суть методу найменших квадратів (МНК).

4.2. Передумови застосування МНК.

4.3. Система нормальних рівнянь.

4.1. Суть методу найменших квадратів

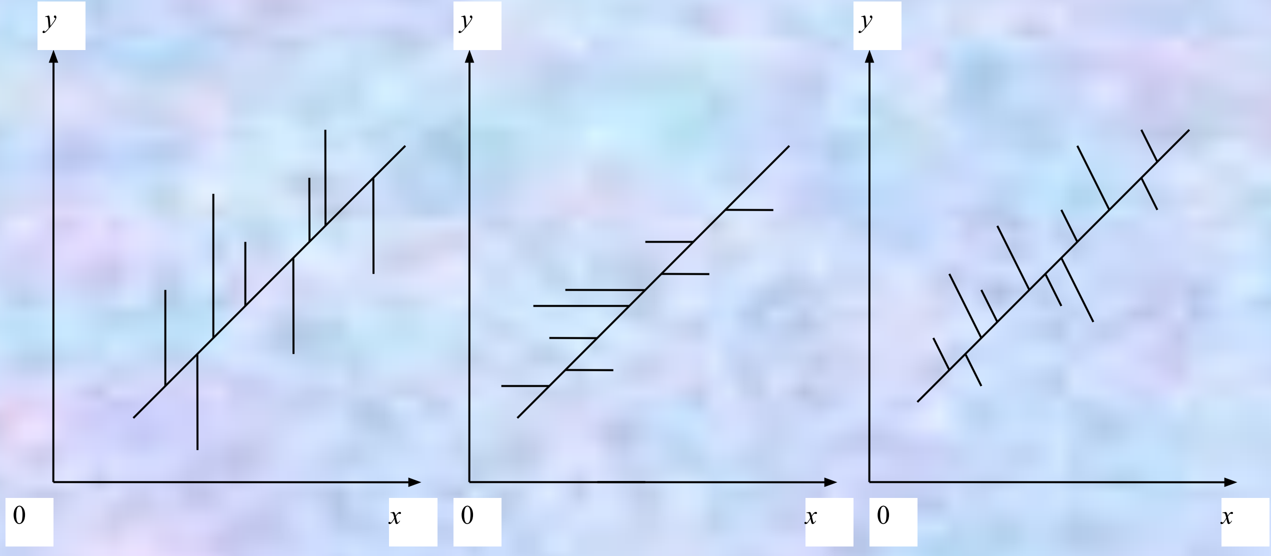


Рис. 4.1. Способи знаходження прямих регресії

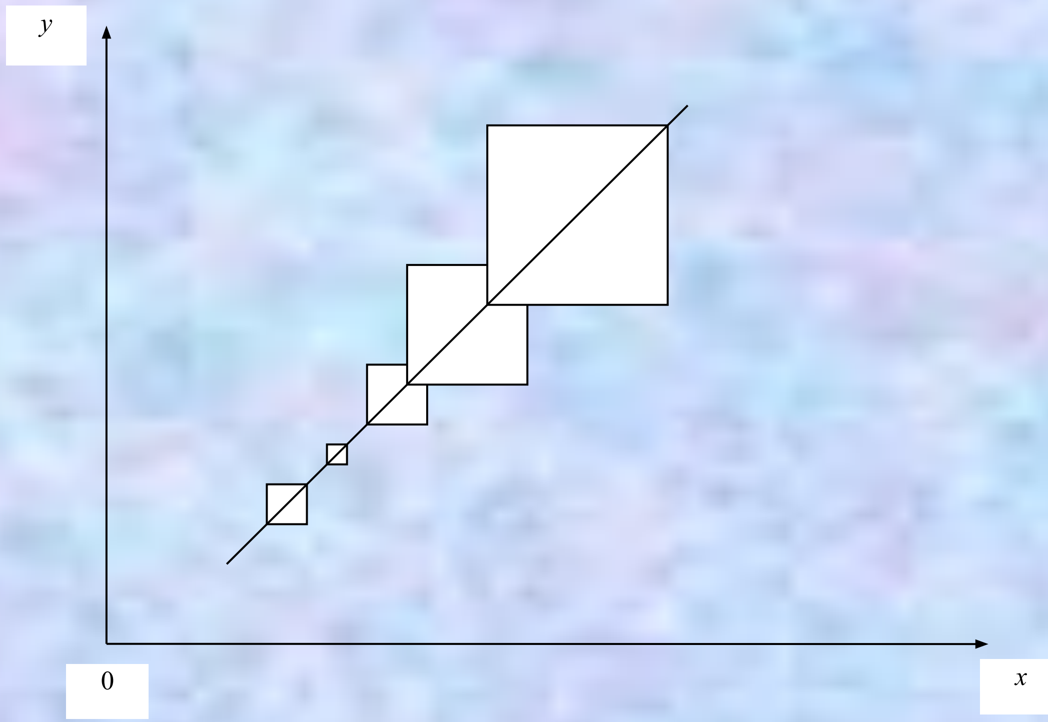


Рис. 4.2. Геометрична інтерпретація методу найменших квадратів

Суть **методу найменших квадратів (МНК)**

полягає **у знаходженні такої <u>теоретичної лінії</u> регресії, яка в порівнянні з іншими проходить <u>найближче до емпіричної лінії регресії</u>,**

тобто дає

найменшу суму квадратів відхилень фактичних **значень результативної ознаки від розрахункових (*теоретичних*) значень**

$$\sum_{i=1}^{n}\left(y_i-\widetilde{y}_i\right)^2 \to \min$$

- **де y – емпіричні (вихідні) дані показника**
- **$\tilde{y}$– теоретичні (розраховані за рівнянням регресії)**

4.2. Передумови застосування МНК

1) Існує лінійний зв'язок між результуючою змінною *y* та факторною змінною *x*, який описується рівняннями регресії

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\tilde{y} = b_0 + b_1 x$$

$i \neq j$

2) Факторна змінна *x* є детерміністичною (невипадковою) величиною.

3) Математичне сподівання (середнє значення) випадкового вектора дорівнює нулю, а дисперсія є невеликою постійною додатньою величиною, яка не залежить від індексу *i*, тобто

$$E\varepsilon_i = 0 \qquad D\varepsilon_i = E\left(\varepsilon_i^2\right) = \sigma^2$$

4) Компоненти вектора є некорельованими випадковими величинами, тобто для кожного $(i, j = 1,2,\ldots,n)$

$$\operatorname{cov}\left(\varepsilon_i, \varepsilon_j\right) = 0$$

5) Часто вважають, що випадкова величина має нормальний закон розподілу з рівним нулю математичним сподіванням і постійною додатньою невеликою дисперсією

$$\varepsilon \sim N\left(0, \sigma^2\right)$$

У даному випадку модель називається класичною нормальною лінійною регресійною моделлю.

- **Зауваження.** У випадку класичної нормальної лінійної регресійної моделі умова 4 еквівалентна умові статистичної незалежності помилок .

4.3. Система нормальних рівнянь

Будемо вважати, що зв'язок між ознаками *x* та *y* є лінійним і описується лінійним рівнянням регресії

$$y = b_0 + b_1 x \qquad (4.3)$$

де *y* – результуюча змінна; b_0, b_1 – параметри рівняння регресії; *x* – факторна змінна; ε– випадкова величина.

У загальному випадку *парна лінійна регресія* є лінійною функцією між залежною змінною *У* і однією пояснюючою змінною *Х*:

$$y = b_0 + b_1 x$$

Це співвідношення називається *теоретичною лінійною регресійною моделлю*

b_0* і *b_1 - теоретичні параметри (теоретичні коефіцієнти) регресії.

x_1	x_2	x_3	$\cdots$	x_{n-1}	x_n
y_1	y_2	y_3	$\cdots$	y_{n-1}	y_n

$$\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} (y_i - y_i) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

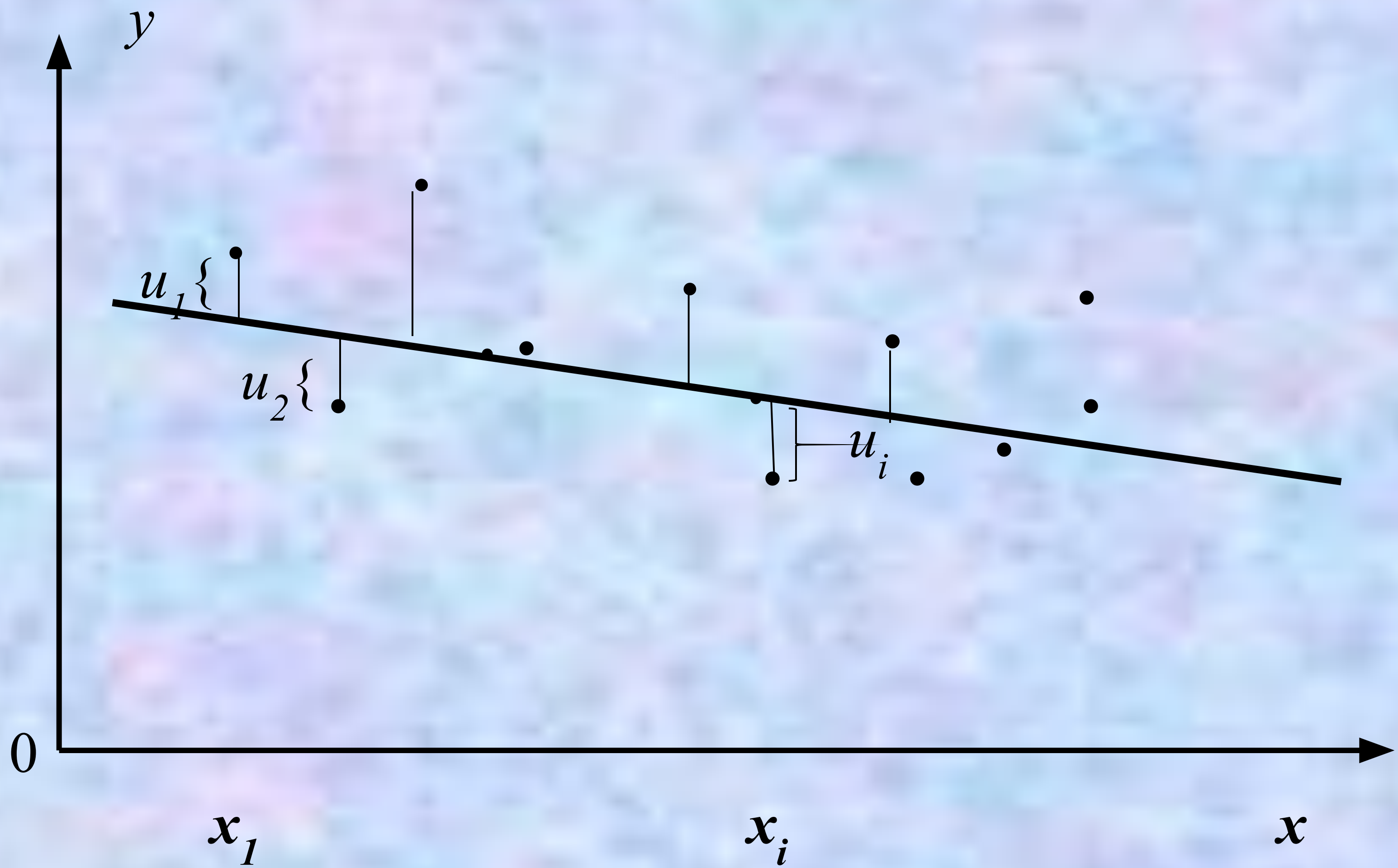
метод найменших модулів (МНМ).

$$\sum_{i=1}^{n} |u_i| = \sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} |y_i - \hat{a}_0 - \hat{a}_1 x_i|$$

$$\sum_{i=1}^{n} u_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

***метод найменших квадратів* (МНК).**

$$Q = \sum \left(y - \tilde{y} \right)^2 \rightarrow \min$$



y
u_1
u_2
u_i
0
x_1
x_i
x

$$\sum_{i=1}^{n} u_i^{\,2} = Q(b_0, b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \to \min$$

Необхідною умовою існування мінімуму неперервно диференційованої функції двох змінних є рівність нулю її частинних похідних.

Так як

$$\frac{\partial y}{\partial b_1} = x; \quad \left(\frac{\partial y}{\partial b_1}\right)_i = x_i;$$

$$\frac{\partial y}{\partial b_0} = 1; \quad \left(\frac{\partial y}{\partial b_0}\right)_i = 1;$$

$$\begin{cases} \dfrac{\partial Q}{\partial b_1} = -2\sum\limits_{i-1}^{n}\left(y_i - b_1 x_i - b_0\right)\cdot x_i = 0 \\ \dfrac{\partial Q}{\partial b_0} = -2\sum\limits_{i-1}^{n}\left(y_i - b_1 x_i - b_0\right) = 0 \end{cases}$$

$$\begin{cases} b_1 \sum\limits_{i-1}^{n} x_i^2 + b_0 \sum\limits_{i-1}^{n} x_i = \sum\limits_{i-1}^{n} x_i y_i; \\ b_1 \sum\limits_{i-1}^{n} x_i + b_0 n = \sum\limits_{i-1}^{n} y_i \end{cases}$$

Система нормальних рівнянь

$$\begin{cases} b_0 n + b_1 \sum x = \sum y \\ b_0 \sum x + b_1 \sum x^2 = \sum xy \end{cases}$$

$$\begin{cases} b_1 \dfrac{\sum\limits_{i-1}^{n} x_i^{\ 2}}{n} + b_0 \dfrac{\sum\limits_{i-1}^{n} x_i}{n} = \dfrac{\sum\limits_{i-1}^{n} x_i y_i;}{n} \\ b_1 \dfrac{\sum\limits_{i-1}^{n} x_i}{n} + b_0 = \dfrac{\sum\limits_{i-1}^{n} y_i;}{n} \end{cases}$$

Позначимо:

$$\bar{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n} \qquad \bar{y} = \frac{\sum\limits_{i=1}^{n} y_i}{n}$$

$$\overline{x^2} = \frac{\sum\limits_{i=1}^{n} x_i^{\ 2}}{n} \qquad \overline{xy} = \frac{\sum\limits_{i=1}^{n} x_i y_i}{n}$$

одержимо

$$\begin{cases} b_1\overline{x^2} + b_0\bar{x} = \overline{xy} \\ b_1\bar{x} + b_0 = \bar{y} \end{cases}$$

звідки маємо

$$\begin{cases} b_1 = \dfrac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} \\ b_0 = \bar{y} - b_1\bar{x} \end{cases}$$

Неважко помітити, що b_1 можна обчислити за формулою:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_x^2}$$

$$S_{xy} = \mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

-вибірковий кореляційний момент випадкових величин X і Y;

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\ 2} - \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

вибіркова дисперсія X

$S_x = \sqrt{S_x^2}$ —стандартне відхилення X.

$$\hat{a}_1 = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x \cdot S_y} \cdot \frac{S_y}{S_x} = r_{xy} \cdot \frac{S_y}{S_x}$$

r_{xy} —вибірковий коефіцієнт кореляції;

S_y —стандартне відхилення Y.

Коефіцієнт кореляції

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}$$

$$0 \le r_{xy} \le 1$$