

*Харківський національний університет
імені В. Н. Каразіна*



**“ІНФОРМАТИКА І ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ
для хіміків”**

Лекція №3 Регресійний аналіз

В. В. Іванов

Кафедра хімічного матеріалознавства

Питання теми

- *Кореляції в хімії;*
- *Уявлення про лінійну регресію;*
- *Метод найменших квадратів;*
- *Дисперсія і коефіцієнт кореляції;*
- *Методики перехресного оцінювання;*
- *Уявлення про робастне оцінювання і
Метод найменших модулів.*

В хімії відомо багато різноманітних кореляцій

Властивість 1	Властивість 2	Аналітична форма кореляції
Атомний радіус r_A	Іонний радіус r_I	$r_A = a_0 + a_1 r_I$
Атомний радіус r_A	Енергія іонізації атомів I	$\lg r_A = a_0 + a_1 \lg I$
Енергія дисоціації E_{diss}	Енергія іонізації молекул, I	$\lg E_{\text{diss}} = a_0 + a_1 \lg I$
Температура плавлення $T_{\text{melt.}}$	Температура кипіння T_{bp}	$T_{\text{melt.}} = a_0 + a_1 T_{\text{bp}}$
Світлопоглинання розчину, оптична густина, D	концентрації поглинаючих частинок [c]	$D = a_1 [c]$
Біоефект C - концентрація речовини $\lg(1/C)$	Ліпофільність $\lg P$	$\lg(1/C) = a_0 + a_1 \lg P + a_1 (\lg P)^2$

№	X	Y
1	X ₁	Y ₁
2	X ₂	Y ₂
3
...
N	X _N	Y _N

Залежна змінна

Незалежна змінна

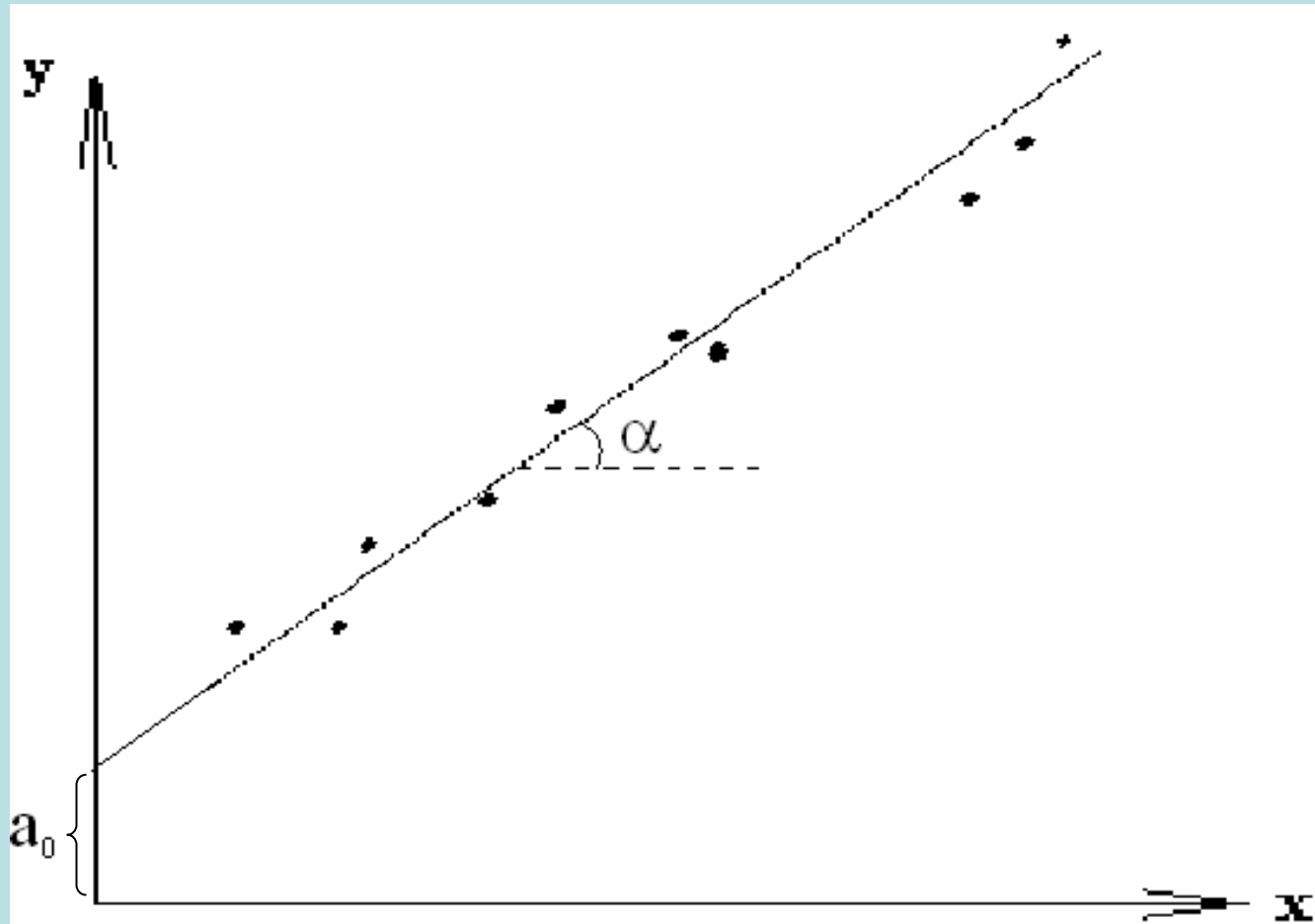
$$y = a_0 + a_1 x$$

Коефіцієнти регресії

Approximation (англ.) - наближення

Геометричний сенс коефіцієнтів регресії

$$y = a_0 + a_1 x$$

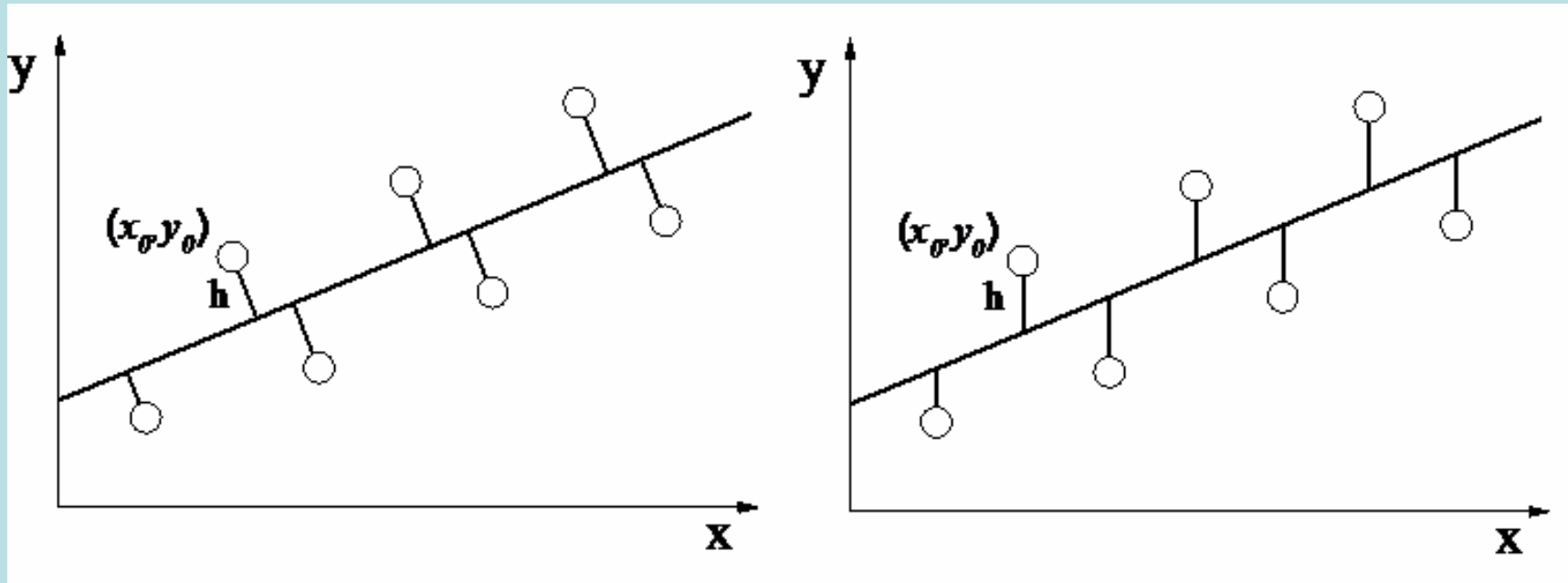


$$a_1 = \operatorname{tg} \alpha$$

Полілінійна регресія (multiple linear regression, MLR)

№	X_1	X_2	...	X_k	Y
1	X_{11}	X_{12}	...	X_{1k}	Y_1
2	X_{21}	X_{22}	...	X_{2k}	Y_2
...
...
N	X_{N1}	X_{N2}	...	X_{Nk}	Y_N

$$y = a_0 + a_1 X_1 + a_2 X_2 + \dots + a_k X_k$$



- Метод ортогональних відстаней;

- Метод найменших модулів; $\Phi = \sum_i |y_i - \hat{y}_i|$

- Метод найменших квадратів (МНК) $F = \sum_i (y_i - \hat{y}_i)^2$

\hat{y}_i - Теоретичне значення

Творці МНК

Карл Фридрих Гаусс (1794)



Адрієн Марі Лежандр (1805)

В багатьох підручниках наводиться такий портрет



Але насправді єдиний портрет що залишився - шарж



Розповсюджений випадок $y = a_0 + a_1 x$

№	x	y
1	x_1	y_1
2	x_2	y_2
3
...
N	x_N	y_N

$$F = F(a_0, a_1) = \sum_i (\hat{y}_i - y_i)^2$$

$$F = F(a_0, a_1) = \sum_i (a_0 + a_1 x_i - y_i)^2$$

$$\begin{cases} \frac{\partial F}{\partial a_0} = 2 \sum_i (a_0 + a_1 x_i - y_i) = 0 \\ \frac{\partial F}{\partial a_1} = 2 \sum_i (a_0 + a_1 x_i - y_i) x_i = 0 \end{cases}$$

$$\begin{cases} \frac{\partial F}{\partial a_0} = 2 \sum_i (a_0 + a_1 x_i - y_i) = 0 \\ \frac{\partial F}{\partial a_1} = 2 \sum_i (a_0 + a_1 x_i - y_i) x_i = 0 \end{cases}$$

$$\begin{cases} \sum_i (a_0 + a_1 x_i - y_i) = 0 \\ \sum_i (a_0 + a_1 x_i - y_i) x_i = 0 \end{cases}$$

$$\begin{cases} a_0 N + a_1 \sum_i x_i - \sum_i y_i = 0 \\ a_0 \sum_i x_i + a_1 \sum_i x_i^2 - \sum_i y_i x_i = 0 \end{cases}$$

$$\begin{cases} a_0 N + a_1 \sum_i x_i - \sum_i y_i = 0 \\ a_0 \sum_i x_i + a_1 \sum_i x_i^2 - \sum_i y_i x_i = 0 \end{cases}$$

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

$$\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i y_i x_i \end{pmatrix}$$

$$y = a_0 + a_1 x$$

$$\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i y_i x_i \end{pmatrix}$$

$$a_0 = \frac{\begin{vmatrix} \sum_i y_i & \sum_i x_i \\ \sum_i y_i x_i & \sum_i x_i^2 \end{vmatrix}}{D} = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i y_i x_i \sum_i x_i}{D}$$

$$a_1 = \frac{\begin{vmatrix} N & \sum_i y_i \\ \sum_i x_i & \sum_i y_i x_i \end{vmatrix}}{D} = \frac{N \sum_i y_i x_i - \sum_i y_i \sum_i x_i}{D}$$

$$D = \det \begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} = N \sum_i x_i^2 - \left(\sum_i x_i \right)^2$$

Приклад №1

x	y
1	0
2	1

$$a_0 = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i y_i x_i \sum_i x_i}{N \sum_i x_i^2 - \left(\sum_i x_i \right)^2}$$

$$a_1 = \frac{N \sum_i y_i x_i - \sum_i y_i \sum_i x_i}{N \sum_i x_i^2 - \left(\sum_i x_i \right)^2}$$

$$\sum_i y_i = 1, \quad \sum_i x_i = 3, \quad \sum_i x_i^2 = 5, \quad \sum_i y_i x_i = 2,$$

$$a_0 = \frac{1 \cdot 5 - 2 \cdot 3}{1} = -1 \quad a_1 = \frac{2 \cdot 2 - 1 \cdot 3}{1} = 1$$

$$y = -1 + x$$

По двух точках - пряма

Приклад № 2 набір з N вимірів $(x_1, x_2, x_3, \dots, x_N,)$

Знайти методом МНК таку величину a_0 , щоб функція $F(a_0)$

$$F = F(a_0) = \sum_i (x_i - a_0)^2$$

була мінімальна

(сума квадратів відстаней від точки a_0 до точок x_i):

$$\frac{\partial F}{\partial a_0} = 2 \sum_i (x_i - a_0) = 0$$

$$\sum_i x_i - Na_0 = 0$$

$$a_0 = \frac{\sum_i x_i}{N} = \bar{x}$$

Середнє значення !

Приклад № 3

Світло поглинання розчину залежить від концентрації поглинаючих світло частинок
(закон Бугера-Ламберта-Бера.)

$$y = a_1 \cdot x$$

$$F(a_1) = \sum_i (y_i - a_1 x_i)^2 \implies \frac{\partial F}{\partial a_1} = -2 \sum_i (y_i - a_1 x_i) x_i = 0$$

$$\sum_i y_i x_i - a_1 \sum_i x_i^2 = 0$$

$$a_1 = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

X(10 ⁻³ моль/л)	Y(поглощения)
0.01	0
1	1
2	3

$$a_1 = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

$$\sum_i x_i^2 \approx 5, \quad \sum_i y_i x_i = 7$$

$$a_1 = \frac{7}{5} = 1.4$$

$$y = 1.4 \cdot x$$

Метод найменших квадратів

$$F = F(a_0, a_1, a_2, \dots, a_k) = \sum_i (\hat{y}_i - y_i)^2$$

$$F = F(a_0, a_1, a_2, \dots, a_k) = \sum_i (a_0 + a_1 x_{i1} + a_2 x_{i2} + \dots + a_k x_{ik} - y_i)^2$$

$$\begin{cases} \frac{\partial F}{\partial a_0} = 0 \\ \frac{\partial F}{\partial a_1} = 0 \\ \dots \\ \frac{\partial F}{\partial a_k} = 0 \end{cases} \quad \begin{cases} \frac{\partial F}{\partial a_0} = 2 \sum_i (a_0 + a_1 x_{i1} + a_2 x_{i2} + \dots + a_k x_{ik} - y_i) = 0 \\ \frac{\partial F}{\partial a_1} = 2 \sum_i (a_0 + a_1 x_{i1} + a_2 x_{i2} + \dots + a_k x_{ik} - y_i) x_{i1} = 0 \\ \dots \\ \frac{\partial F}{\partial a_k} = 2 \sum_i (a_0 + a_1 x_{i1} + a_2 x_{i2} + \dots + a_k x_{ik} - y_i) x_{ik} = 0 \end{cases}$$

Метод найменших квадратів

$$\left\{ \begin{array}{l} a_0 N + a_1 \sum_i x_{i1} + a_2 \sum_i x_{i2} + \dots + a_k \sum_i x_{ik} = \sum_i y_i \\ a_0 \sum_i x_{i1} + a_1 \sum_i x_{i1} x_{i1} + a_2 \sum_i x_{i2} x_{i1} + \dots + a_k \sum_i x_{ik} x_{i1} = \sum_i y_i x_{i1} \\ \dots \\ a_0 \sum_i x_{ik} + a_1 \sum_i x_{ik} x_{i1} + a_2 \sum_i x_{i2} x_{ik} + \dots + a_k \sum_i x_{ik} x_{ik} = \sum_i y_i x_{ik} \end{array} \right.$$

$$\begin{pmatrix} N & \sum_i x_{i1} & \dots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \dots & \sum_i x_{ik} x_{i1} \\ \dots & \dots & \dots & \dots \\ \sum_i x_{ik} & \sum_i x_{i1} x_{ik} & \dots & \sum_i x_{ik}^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i y_i x_{i1} \\ \dots \\ \sum_i y_i x_{ik} \end{pmatrix}$$

$$Aa = B$$

Метод наименших квадратов

$$\begin{pmatrix} N & \sum_i x_{i1} & \dots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \dots & \sum_i x_{ik} x_{i1} \\ \dots & \dots & \dots & \dots \\ \sum_i x_{ik} & \sum_i x_{i1} x_{ik} & \dots & \sum_i x_{ik}^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i y_i x_{i1} \\ \dots \\ \sum_i y_i x_{ik} \end{pmatrix}$$

Предполагаем, что матрица невырождена

$$\begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{pmatrix} = \begin{pmatrix} N & \sum_i x_{i1} & \dots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \dots & \sum_i x_{ik} x_{i1} \\ \dots & \dots & \dots & \dots \\ \sum_i x_{ik} & \sum_i x_{i1} x_{ik} & \dots & \sum_i x_{ik}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_i y_i \\ \sum_i y_i x_{i1} \\ \dots \\ \sum_i y_i x_{ik} \end{pmatrix}$$

Нелінійні залежності (лінеаризуємі)

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \dots + a_k x_i^k$$

$$X_1 = X, X_2 = X^2, \dots, X_k = X^k$$

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \dots + a_k x_{ik}$$

$$y_i = a_0 \cdot e^{a_1 x_i}$$

$$\ln y_i = \ln a_0 + a_1 x_i$$

Задача лінійна в
координатах

$$\ln y_i, x_i$$

$$y_i = a_0 \cdot x_i^{a_1}$$

$$\ln y_i = \ln a_0 + a_1 \ln x_i$$

Задача лінійна в
координатах

$$\ln y_i, \ln x_i$$

Особливість задачі МНК

$$F = \sum_i (y_i - \hat{y}_i)^2$$

$$y = ax + b$$

~~$$x = (y - b) / a$$~~

Повернемося до прикладу №3 (слайд 15)

$$x = b_1 y$$

$$b_1 = \frac{\sum_i y_i x_i}{\sum_i y_i^2}$$

Знаходимо b_1

$$\sum_i y_i^2 = 1 \cdot 1 + 3 \cdot 3 = 10 \quad b_1 = \frac{7}{10} = \underline{0.7}$$

З першого рівняння регресії (див. слайд 16)

$$y = 1.4 \cdot x \quad b_1 \approx 1/a_1 = 1/1.4 = \underline{0.7143}$$

розрізняються !

Міри адекватності отриманих рівнянь

$$\sigma^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{N - p}$$

$p = k + 1$ – кількість параметрів

$$SD = \sigma = \sqrt{\sigma^2} \quad \text{Standard Deviation}$$

У випадку: $N = 2$ $p = 2$

$$\sigma^2 = \frac{0}{0} \quad \text{дисперсія не визначена}$$

До прикладу № 3

$$y = 1.4 \cdot x$$

x (10^{-3} моль/л)	Y (погл.) експ	\hat{y} (погл.) теор.
0.01	0	0.014
1	1	1.4
2	3	2.8

$$\begin{aligned}\sigma^2 &= \frac{1}{3-1} \sum_i (y_i - \hat{y}_i)^2 = \frac{1}{2} \left((1-1.4)^2 + (3-2.8)^2 \right) = \\ &= \frac{1}{2} (0.4^2 + 0.2^2) = \frac{1}{2} (0.16 + 0.04) = 0.1\end{aligned}$$

$$SD = \sigma \approx 0.32 \quad \text{Точність} = \pm 0.32$$

коефіцієнт кореляції (множинний коефіцієнт кореляції)

Наскільки узгоджуються зміни величин x і y ?

$$y = \{y_1, y_2, \dots, y_N\} \quad x = \{x_1, x_2, \dots, x_N\}$$

$$\bar{y} = \frac{\sum y_i}{N}$$

$$\bar{x} = \frac{\sum x_i}{N}$$

Центрування:

$$y_i - \bar{y}$$

$$x_i - \bar{x}$$

Нормування:

$$y_i = \frac{y_i - \bar{y}}{\sqrt{\sum_i (y_i - \bar{y})^2}}$$

$$x_i = \frac{x_i - \bar{x}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

$$\sum_i y_i^2 = 1$$

$$\sum_i x_i^2 = 1$$

коефіцієнт кореляції (множинний коефіцієнт кореляції)

$$y_i \rightarrow \frac{y_i - \bar{y}}{\sqrt{\sum_i (y_i - \bar{y})^2}} \quad x_i \rightarrow \frac{x_i - \bar{x}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Скалярний добуток двох нормованих (і центрованих) векторів.

$$r_{y,x} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (y_i - \bar{y})^2 (x_i - \bar{x})^2}}$$

$$0 \leq |r| \leq 1$$

“відмінні”

$$|r| > 0.99$$

“добрі”

$$0.98 \leq |r| \leq 0.99$$

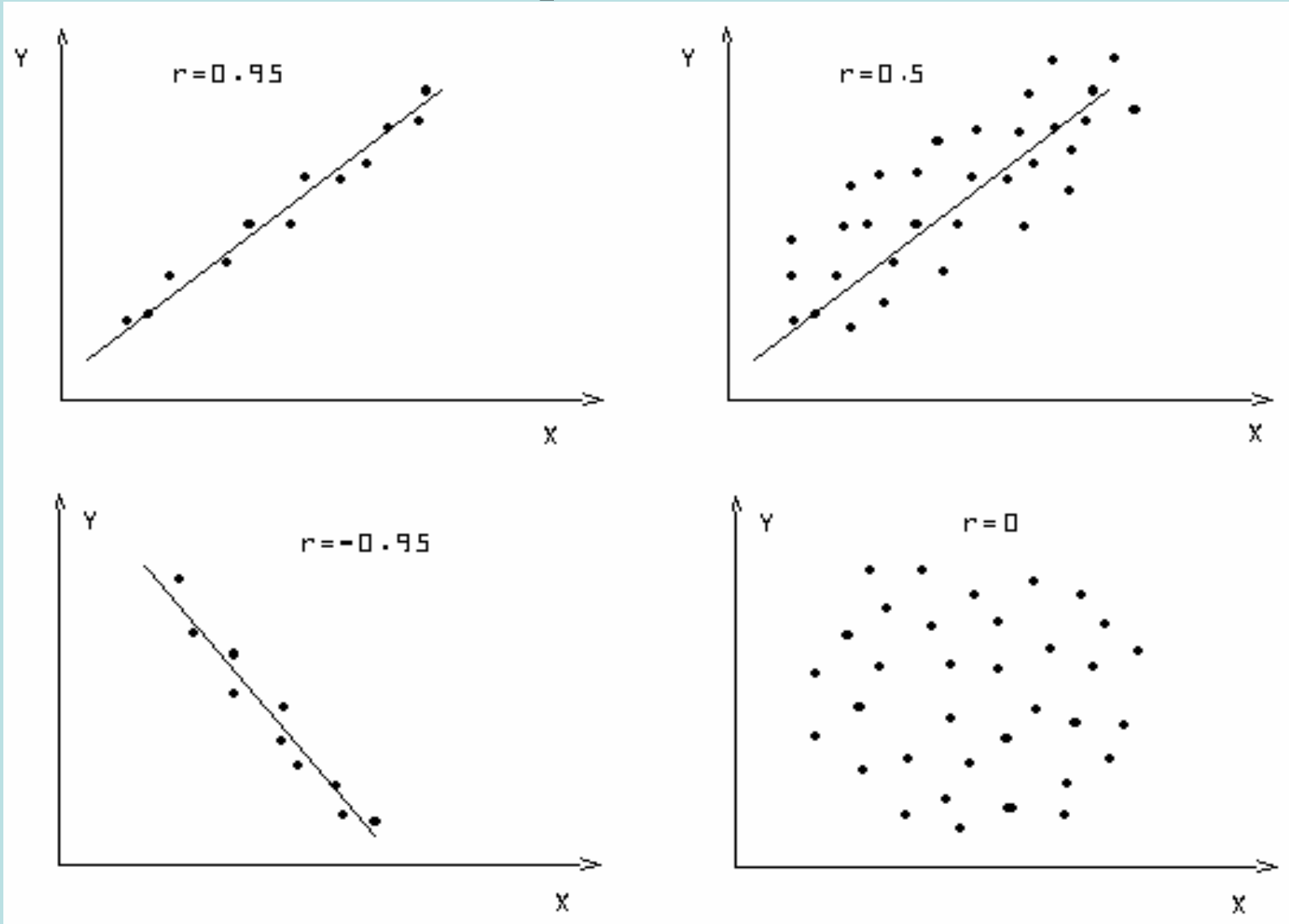
“задовільні”

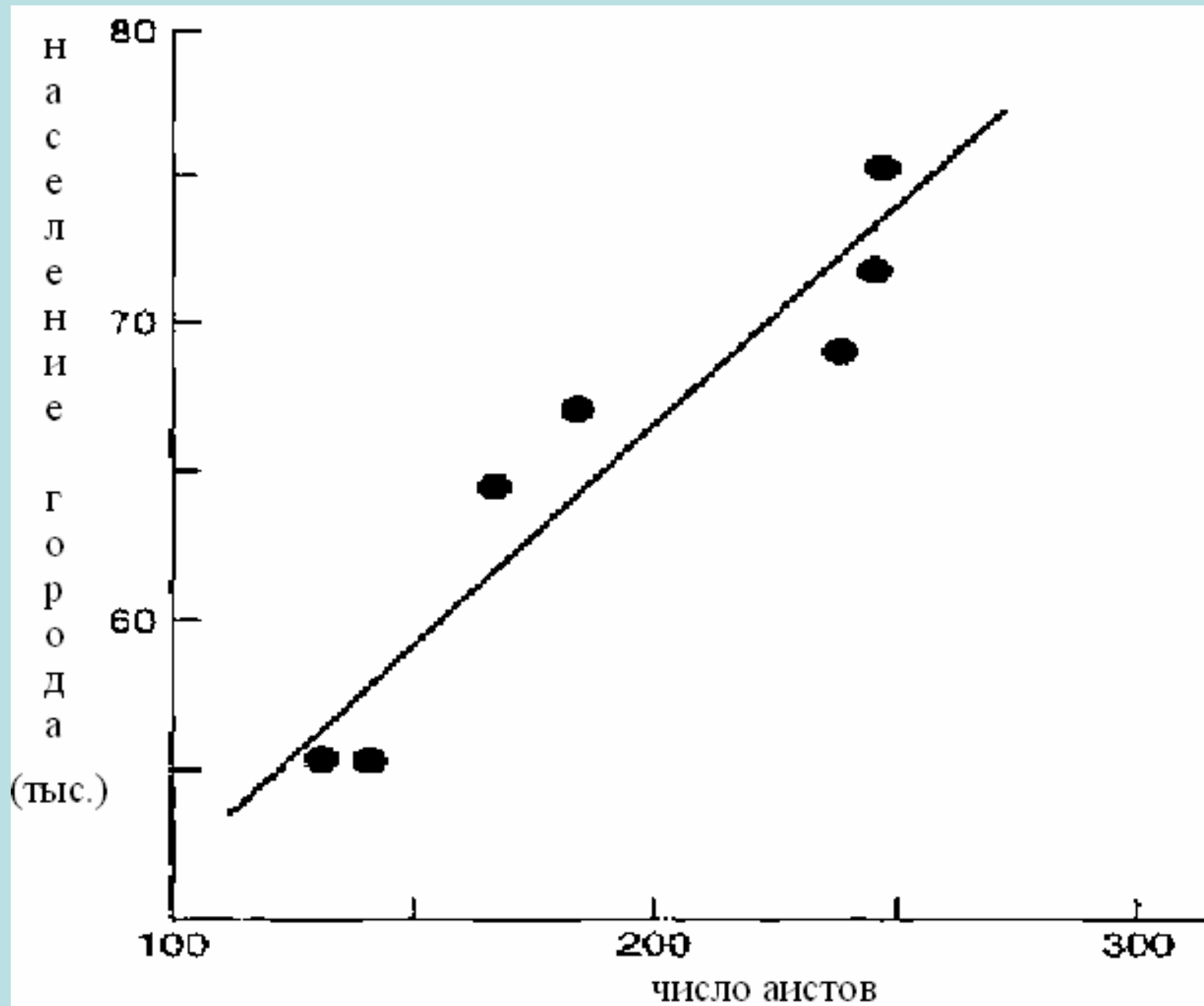
$$0.95 \leq |r| < 0.98$$

“погані”

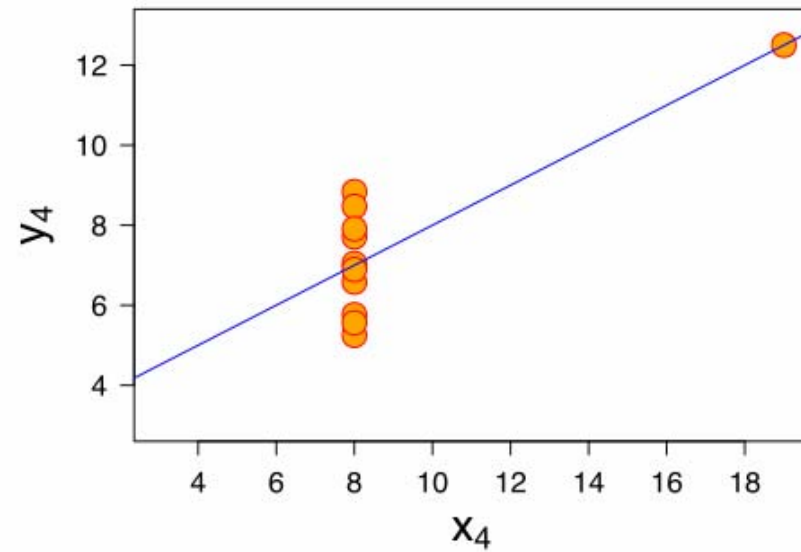
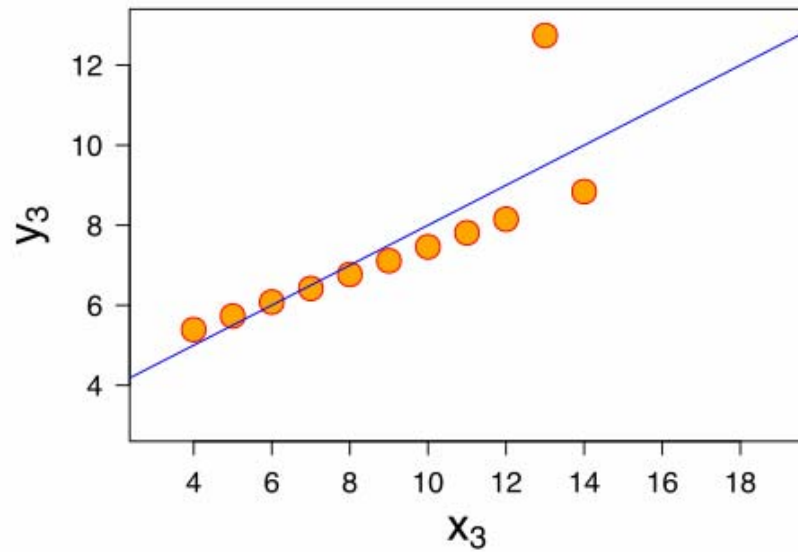
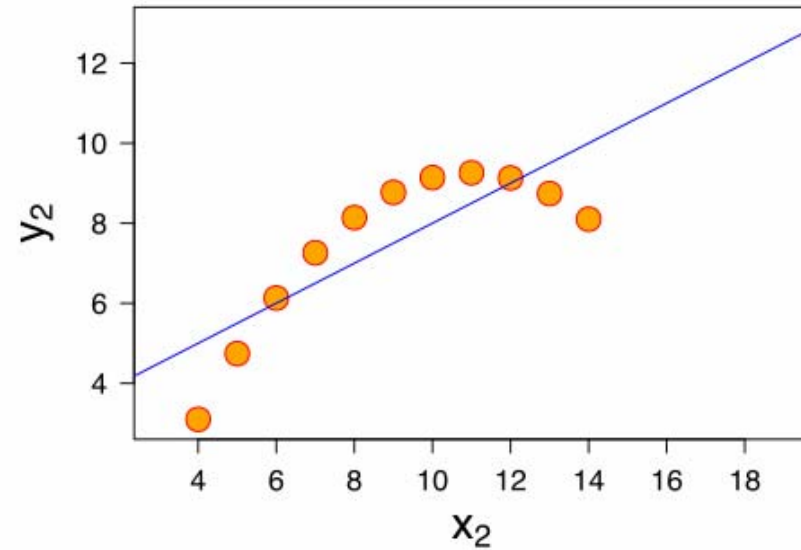
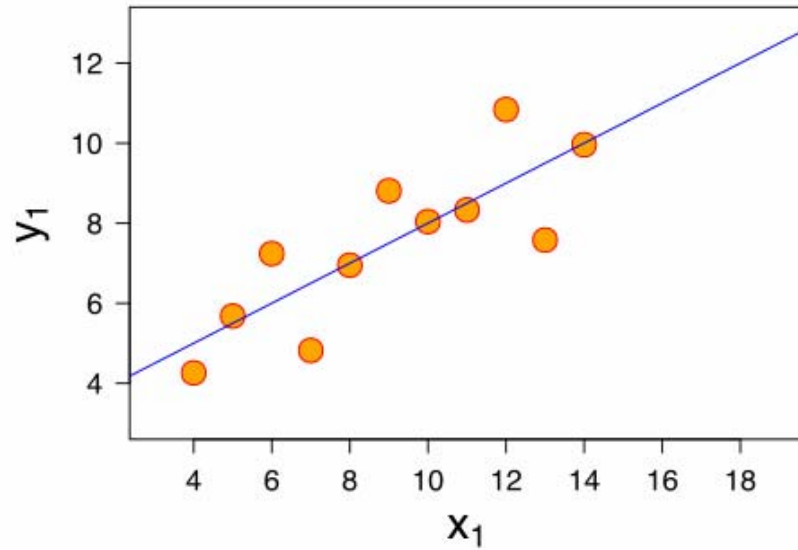
$$|r| < 0.9$$

Приклади





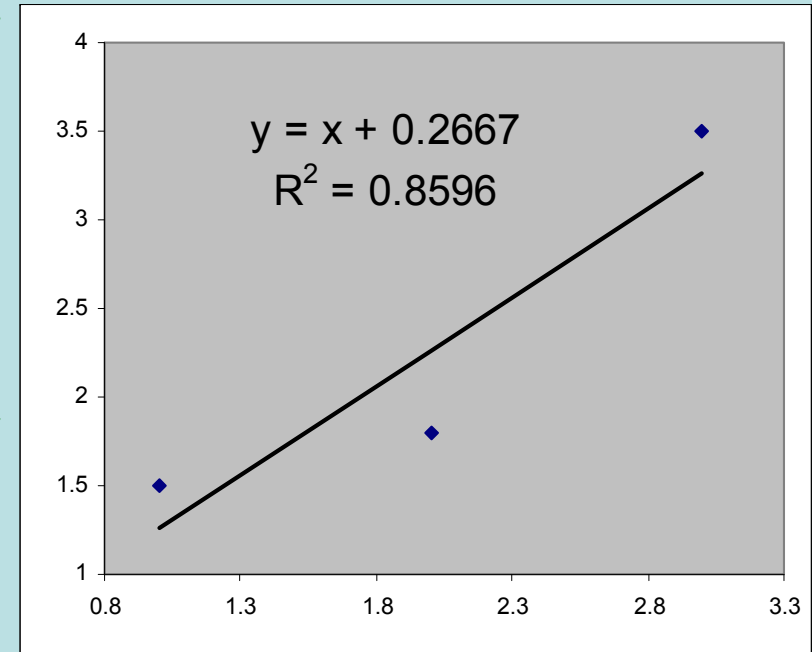
Abscomb Plots



Процедура LOO (Leave-one-out cross-validation)

Перехрестне оцінювання

X	Y	Y(calc)	eqs.	Y(pred)
1	1.5	1.267	$Y=1.7x-1.6$	0.1
2	1.8	2.267	$Y=x+0.5$	2.5
3	3.5	3.267	$Y=0.3x+1.2$	2.1



Calculated

$$\sigma^2 = \frac{1}{3-2} \sum_i (y_i - \hat{y}_i)^2 =$$

$$\left((1.5 - 1.267)^2 + (1.8 - 2.267)^2 + (3.5 - 3.267)^2 \right)$$

$$\approx 0.33 \quad \underline{\sigma \approx \sqrt{0.33} = 0.57}$$

Predicted

$$\sigma^2 = 4.41 \quad \underline{\sigma \approx \sqrt{4.41} = 2.1}$$

Щодо вибору апроксимуючої Функції

$$\left. \begin{array}{l} y = ax + b \\ \frac{dy}{dx} = a \end{array} \right\} \longrightarrow \begin{array}{l} \frac{\Delta y}{\Delta x} = \text{const} \\ \Delta y = y_i - y_{i-1} \\ \Delta x = x_i - x_{i-1} \end{array}$$

$y = a_0 x^{a_1}$	$\frac{\Delta \ln y}{\Delta \ln x} = \text{const}$
$y = a_0 a_1^x$	$\frac{\Delta \ln y}{\Delta x} = \text{const}$
$y = a_0 + a_1 x + a_2 x^2$	$\frac{\Delta y^2}{\Delta x^2} = \text{const}$
$y = \frac{x}{a_0 + a_1 x}$	$\frac{\Delta \left(\frac{x}{y} \right)}{\Delta x} = \text{const}$

Опис аналітично заданих функцій

$$\text{STO} = \text{N GTO} \quad e^{-\alpha r} \approx \sum_i C_i e^{-\zeta_i r^2}$$

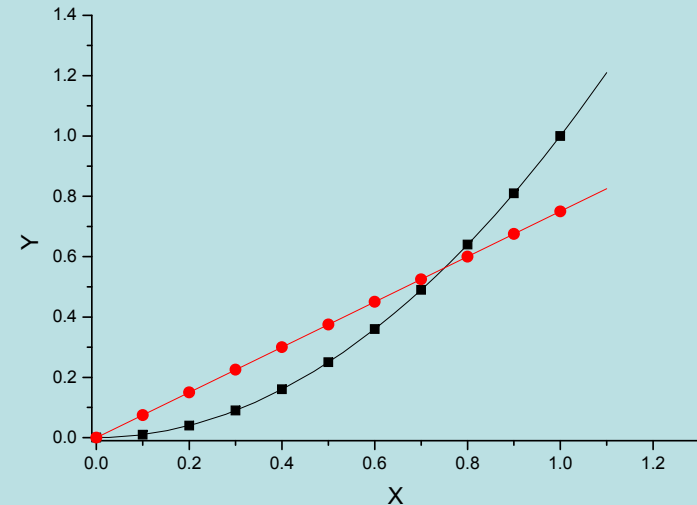
$$J = \int_0^{\infty} \left(e^{-\alpha r} - \sum_i C_i e^{-\zeta_i r^2} \right)^2 dr$$

Приклад № 4

$$y(x) = x^2 \quad x = [0 \div 1]$$

$$f(x) = a_1 x$$

$$F(a_1) = \int_0^1 (x^2 - a_1 x)^2 dx$$



$$F(a_1) = \int_0^1 (x^2 - a_1 x)^2 dx$$

$$F(a_1) = \int_0^1 (x^4 - 2a_1 x^3 + a_1^2 x^2) dx = \int_0^1 x^4 dx - 2a_1 \int_0^1 x^3 dx + a_1^2 \int_0^1 x^2 dx$$

$$F(a_1) = \frac{x^5}{5} \Big|_0^1 - 2a_1 \frac{x^4}{4} \Big|_0^1 + a_1^2 \frac{x^3}{3} \Big|_0^1 = \frac{1}{5} - a_1 \frac{1}{2} + a_1^2 \frac{1}{3}$$

$$\frac{dF(a_1)}{da_1} = -\frac{1}{2} + a_1 \frac{2}{3} = 0 \qquad a_1 = \frac{3}{4} = 0.75$$

$$y(x) = 0.75x$$

Наближена апроксимація параболи лінійною функцією

X	$y = x^2$
0	0
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25
0.6	0.36
0.7	0.49
0.8	0.64
0.9	0.81
1	1

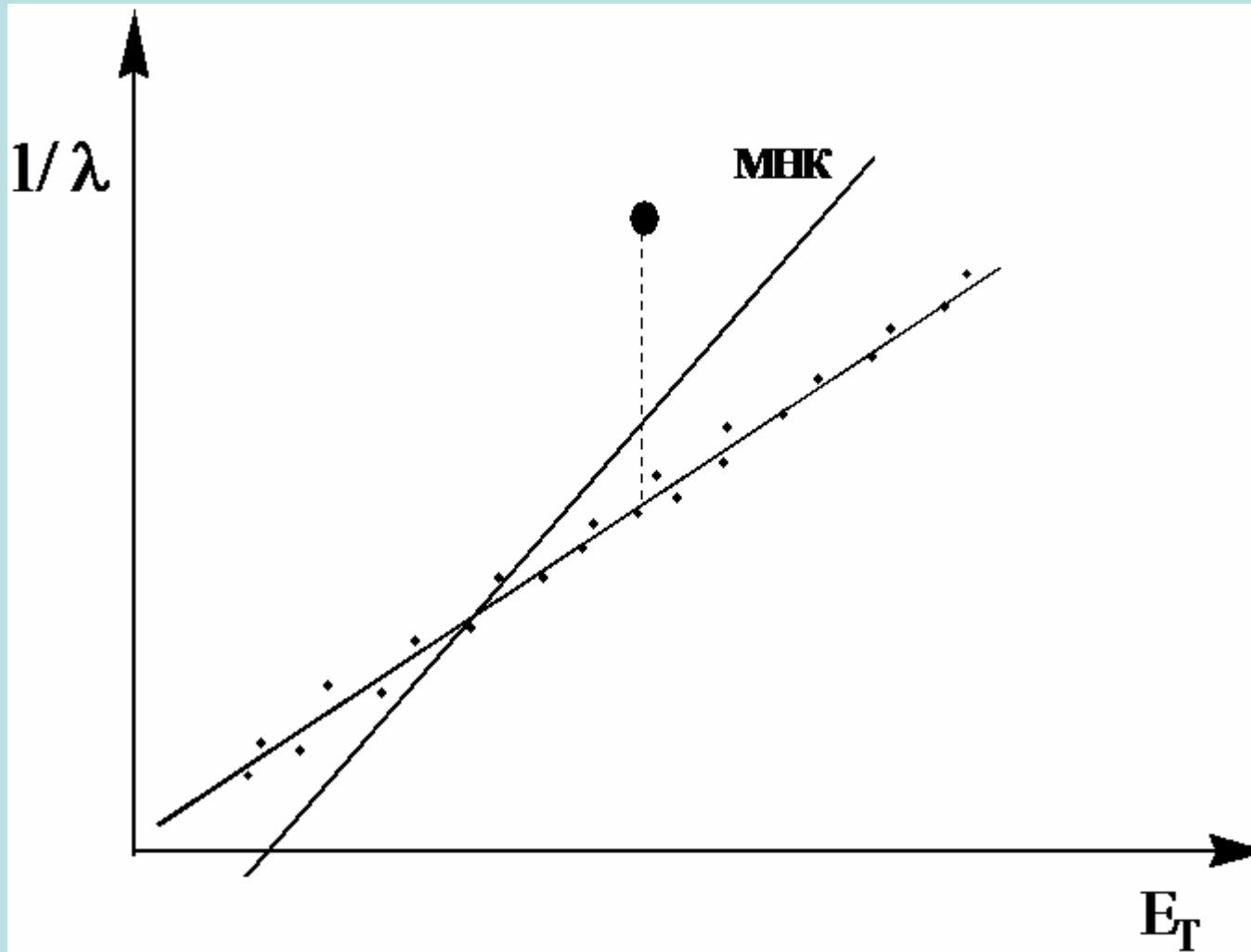
$$a_1 = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

$$a_1 = \frac{\sum_i x_i^3}{\sum_i x_i^2}$$

$$a_1 \approx 0.7857$$

Нагадаємо точне значення: $a_1 = \frac{3}{4} = 0.75$

Недоліки МНК (не робастність)



Зважений метод найменших квадратів

$$F = F(a_0, a_1, a_2, \dots, a_k) = \sum_i (y_i - \hat{y}_i)^2 \quad \text{МНК}$$

$$F = F(a_0, a_1, a_2, \dots, a_k) = \sum_i p_i (y_i - \hat{y}_i)^2 \quad \text{Зважений МНК}$$

$$p_i = \frac{1}{\sigma_i^2}$$

Перелік посилань

1. Поллард Дж., Справочник по вычислительным методам статистики, М.Финансы и статистика, 1982, 344 с.
2. Лоусон Ч., Хенсон Р., Численное решение задач метода наименьших квадратов, М."Наука", 1986, 230 с.
3. Демиденко Е.З. Линейная и нелинейная регрессии, М. Финансы и статистика, 1981, 302 с.
4. А.И.Лесникович, С.В.Левчик, Корреляции в современной химии, Минск, Университетское, 1989
5. Н.Ф.Степнов, М.Е.Ерлыкина, Г.Г.Филиппов, Методы линейной алгебры в физической химии, из-во московского университета, 1976, 360 с.
6. Мудров В.И., Кушко В.Л. Метод наименьших модулей, М.Знание, 1971, 59 с.
7. Мудров В.И., Кушко В.Л. Методы обработки измерений, М.Советское радио, 1976, 190 с.

МНМ

«Истинное знание заключается в том, чтобы знать почему вещи такие какие они есть, а не только в том какие они»

сэр Исая Бёрлин.

*([6 июня 1909](#), [Рига](#) — [5 ноября 1997](#), [Оксфорд](#)) **Английский философ***